# CrowdWater: Motivations of Citizen Scientists, the Accuracy and the Potential of Crowd-Based Data for Hydrological Model Calibration

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von


Simon Etter

aus

Winterthur ZH


Promotionskommission

Prof. Dr. Jan Seibert (Vorsitz)

Dr. Ilja van Meerveld

Prof. Dr. Kai Niebert


Zürich, 2020

# Abstract

Citizen science is a promising tool for the collection of environmental data because it allows data to be collected at many more locations than individual scientists could cover. The citizen science project CrowdWater aims to collect hydrological data using a smartphone app but does not require any physical installations in the stream or the ground. With the app, citizens can collect water level class data using a virtual staff gauge, submit streamflow estimates, report a qualitative soil moisture class, or report the state of intermittent streams or plastic pollution. This thesis focuses on the water level class and streamflow estimates.

I investigated the motivations of the citizen scientists that contributed to CrowdWater and compared it to citizen scientists who contributed to the Naturkalender project using an online questionnaire. Naturkalender is an Austrian citizen science project that uses a similar app as CrowdWater and focuses on the collection of phenological observations of indicator plant and animal species. Citizen scientists who contribute to the projects are mainly driven by their desire to contribute to science, help society and to protect the environment, as well as to learn something new. While most CrowdWater participants agreed that their motivations to engage in the project are also fulfilled by participation, most Naturkalender participants agreed that enjoyment and learning something new were also being fulfilled by their participation. While the enjoyment aspect was not a major reason to join the projects, it was a main reason to continue contributing to both projects. This is encouraging for the further collection of crowd-based water level class observations.

The quality of crowd-based streamflow and water level class observations were first assessed in a survey along nine streams in Switzerland. The results showed that water level classes were easier to estimate and had fewer and smaller errors than the streamflow estimates. The quality of the crowd-based water level class observations obtained with the CrowdWater app was also assessed by comparing them to measured water levels. The correlation between the water level class observation and the water level measurements was very good when the staff was gauge well placed. The correlation was better when the observations were made by individual citizen scientists using the app, rather than multiple citizen scientists who were asked to contribute using signs. Some of the dedicated citizen scientists contributed more than one observation per week.

A modelling study, using synthetic streamflow time series based on the errors from the survey showed that these data are not useful for calibrating the hydrological model HBV-light because the errors are too large. Model calibration with synthetic water level class time series based on errors from the survey, however, showed that these data are valuable because they led to a significantly better model performance compared to simulations using random parameter sets that represent a situation without any data. The model performance was little affected by errors or the number of water level classes that were used but depended on the number of observations and the timing of the observations throughout the year.

This thesis thus shows that citizens are willing to participate in hydrological data

collection, that the quality of these data are good and that these data are useful for the calibration of hydrological models. Therefore, crowd-based water level class observations are a promising source of data for catchments where otherwise no information or very little information on streamflow is available. These data could potentially be used for the calibration of models that can be used for flood warning or to predict the effects of droughts.

# Plain Language Summary

Data and information on the amount of water in streams are important for the management of our water resources. Streamflow data can be used to predict floods or help to regulate the withdrawal of water from rivers during dry periods. Because the continuous collection of this data is associated with considerable cost and effort, such data are often not up-to-date or not available at all for many regions around the world. In addition, the global number of active monitoring stations has decreased in recent years.

One way to collect data in regions where no data is otherwise available is citizen science. The citizen scientists in the CrowdWater project use a smartphone app to collect data on water levels in rivers. The information is read from a virtual yardstick with water level classes in combination with a photo of the river.

The greatest motivation for many of the citizen scientists to participate, was the hoped-for contribution to research. Other important motivators were to contribute to environmental protection, to learn something new and to help society. Not all of these motivations were fulfilled by participating for all the citizen scientists surveyed, but many stated that they enjoyed participating and that by participating they acted according to their values and beliefs.

Surveys of passers-by showed that it is very difficult for citizen scientists to estimate the streamflow directly or via an estimated width, average depth and flow velocity of a river. Estimating water level classes on the basis of the virtual yardstick proved to be easier and the streamflow quantities calculated from it were more accurate.

The comparison of time series of water level class estimates from citizen scientists who contributed either via the CrowdWater app or with forms deposited on fixed mailboxes, showed that the data collected with the app is of higher quality. This was due to the larger number of individuals who contributed for one location, while the majority of contributions to a time series in the app were made by a single contributor.

The streamflow estimates from the passers-by surveys were subject to very large uncertainties and therefore proved to be too imprecise for the calibration of hydrological models. The water level class observations, on the other hand, proved to be potentially useful to calibrate hydrological models when no other measured discharge data are available. These results show that the water level class estimation approach has the potential to generate valuable data where no other data are available and thereby to improve the management of water resources in such regions.

# Zusammenfassung


Daten und Informationen über Fliessgewässer sind wichtig für die Verwaltung unserer Wasserressourcen. So ermöglichen beispielsweise Abflussdaten aus Flüssen die Vorhersage von Hochwassern oder helfen, die Entnahme von Wasser aus Flüssen während Trockenperioden sinnvoll zu regulieren. Weil die kontinuierliche Erfassung dieser Daten mit erheblichen Kosten und Aufwand verbunden ist, sind solche Daten vielerorts auf der Welt nicht aktuell oder gar nicht verfügbar. Zudem nahm die globale Anzahl der aktiven Messstellen in den letzten Jahren ab.

Eine Möglichkeit, um Daten in Regionen zu sammeln, wo sonst keine Daten vorhanden sind, ist der Ansatz der Citizen Science (deutsch Bürgerwissenschaften). Dieser Ansatz setzt auf den Miteinbezug von Privatpersonen in die Forschung. Die Citizen Scientists im Projekt CrowdWater sammeln mittels einer Smartphone App Daten zum Wasserstand in Flüssen. Die Informationen werden in Klassen von einer virtuellen Messlatte auf einem Foto des Flusses von den Citizen Scientists abgelesen und mit einem neuen Foto hochgeladen.

Den erhofften Beitrag, den die Citizen Scientists mit ihrer Teilnahme zur Forschung leisten konnten, war für viele die grösste Motivation mitzumachen. Weitere wichtige Motivatoren waren, einen Beitrag zum Umweltschutz zu leisten, etwas Neues zu lernen und der Gesellschaft zu helfen. Durch die Teilnahme wurden nicht alle diese Motivationen bei allen befragten Citizen Scientists erfüllt, jedoch gaben viele an, dass sie Spass bei der Teilnahme haben und dass sie mit ihrer Teilnahme entsprechend ihrer Überzeugungen handeln.

Befragungen von Passantinnen und Passanten zeigten, dass es für Citizen Scientists sehr schwer ist, den Abfluss direkt oder via Schätzungen der Breite, der mittlere Tiefe und der Fliessgeschwindigkeit eines Flusses zu bestimmen. Das Schätzen von Wasserstandsklassen anhand der virtuellen Messlatte erwies sich als einfacher und die daraus errechneten Abflussmengen als genauer.

Der Vergleich von Zeitreihen von Wasserstandsklassen-Beobachtungen von Citzen Scientists, die mit der CrowdWater App oder mit Formularen an fix installierten Briefkästen beitrugen, zeigte eine bessere Datenqualität der mit der App gesammelten Daten. Grund hierfür war die grössere Anzahl an Einzelpersonen, die mittels Formularen an einer Stelle schätzten, während in der App mehrheitlich dieselbe Person Beobachtungen einer Stelle machte.

Die Abflussschätzungen aus den Befragungen der Passanten waren mit sehr grossen Unsicherheiten behaftet und erwiesen sich deshalb als zu ungenau für die Kalibration von hydrologischen Modellen. Die Beobachtungen der Wasserstandsklassen hingegen er-

wiesen sich als potenziell nützlich, um hydrologische Modelle zu kalibrieren, wenn sonst keine gemessenen Abflussdaten vorhanden sind. Diese Ergebnisse zeigen, dass der Ansatz der Wasserstands-Klassen-Beobachtungen das Potential, hat wertvolle Daten zu generieren, wo sonst keine Daten vorhanden sind. Damit wird eine bessere Verwaltung von Wasserressourcen auch in solchen Regionen ermöglicht.

# Papers and Author Contributions

## List of papers

**Paper I** Seibert, J., Strobl, B., Etter, S., Hummer, P. and van Meerveld, H. J.: Virtual Staff Gauges for Crowd-Based Stream Level Observations, *Frontiers in Earth Science*, 7, doi:10.3389/feart.2019.00070, 2019.

**Paper II** Etter, S., van Meerveld, H. J., Seibert, J., Strobl, B. and Niebert, K.: What motivates people to participate in environmental citizen science projects?, *Citizen Science: Theory and Practice*, resubmitted after moderate revisions.

**Paper III** Strobl, B., Etter, S., van Meerveld, H. J., Seibert, J., Strobl, B. and Etter, S.: Accuracy of crowd-based streamflow and stream level class estimates, *Hydrological Sciences Journal*, 1–19, doi:10.1080/02626667.2019.1578966, 2019.

**Paper IV** Etter, S. Strobl, B., van Meerveld, H.J. and Seibert J.: Accuracy of crowd-based water level classes. *Hydrological Processes*, being revised.

**Paper V** Etter, S., Strobl, B., Seibert, J. and van Meerveld, H. J.: Value of uncertain streamflow observations for hydrological modelling, *Hydrology and Earth System Sciences*, 22, 5243–5257, doi:10.5194/hess-22-5243-2018, 2018.

**Paper VI** Etter, S., Strobl, B., van Meerveld, H. J. (Ilja) and Seibert, J.: Value of crowd-based water level class observations for hydrological model calibration, *Water Resources Research*, doi:10.1029/2019WR02610, 2020.

## Author contributions

**Paper I:** This paper was mainly written by Jan Seibert. Barbara Strobl, Ilja van Meerveld and I helped to shape the study based on our experience with the CrowdWater project. Barbara Strobl and I provided the graphics and user statistics of the CrowdWater app. Philipp Hummer wrote the section about the app design and all co-authors provided comments on the draft manuscripts.

**Paper II:** This paper was my idea. I went to a workshop on motivation in citizen science in March 2018, where I received the necessary information and knowledge to create the questionnaire. I developed the questionnaire with inputs from all co-authors, conducted the survey, analysed the results and wrote the first draft of the manuscript. Barbara Strobl helped in translating the questionnaire into German and Kai Niebert provided valuable expertise in the selection of statements for the questionnaire. All co-authors provided help in shaping the study and selecting the relevant findings during the writing phase. I wrote the first draft of the manuscript and created all figures. All co-authors contributed to the editing of the manuscript.

**Paper III:** For this paper, I contributed mainly in the data collection. Together with Barbara Strobl, I spent several Saturdays and Sundays near Swiss streams to collect the estimates of the water level classes and streamflow from people who passed by the stream. The paper was mainly written by Barbara Strobl. Like all co-authors, I provided comments on the manuscript.

**Paper IV:** I had the lead in designing the study based on valuable comments and suggestions by all co-authors. Barbara Strobl helped in gathering the official water level data from Austrian agencies. Barbara Strobl and I conducted the necessary field work. All the crowd-based data were collected by multiple persons in the app and at our field stations during the years 2016-2019. I wrote the manuscript with valuable inputs from all co-authors.

**Papers V and VI:** I had the lead in designing the studies, creating the synthetic data sets, analysing the results and writing the manuscript. The co-authors helped to shape the studies with constructive comments and ideas and provided feedback on the draft manuscripts.

# Contents

# 1
# Introduction

## 1.1 Importance of hydrological data

Hydrological data contain valuable information on water resources and their variations in the annual cycle. Streamflow records, for example, can be used to quantify current surface water resources and changes over time. Hydrological data are also crucial for the management of water resources (Gilbert, 2010), e.g. to allocate water resources for a growing population (Paper I), to avoid uncontrolled release of wastewater (Davids et al., 2018), optimize water releases for hydropower production (Kundzewicz, 1997), ensure sufficient water for cooling of nuclear power plants (Kirkwood, 1982), sustainable water withdrawals for agriculture and other industrial uses, as well as for planning flood or drought protection and prevention measures (Buytaert et al., 2014). It is useful for scientific and planning purposes to compare historic and recent data to identify systematic shifts and trends in hydrological processes (Milly et al., 2015; Kundzewicz, 2004) and to ultimately allow the modelling of future changes (Hannah et al., 2011). In many regions of the world, however, hydrological and meteorological instruments to obtain these data are scarcely deployed or maintained (Hannah et al., 2011; Sivapalan, 2003). Areas where data are lacking are often also the areas that are most vulnerable to extreme hydrological conditions and events (Walker et al., 2016). The lack of available data, furthermore, results in uncertain predictions about global trends in streamflow, and the occurrence of floods and drought (Stocker et al., 2013). This is illustrated for the availability of streamflow gauge data from the global runoff database in 1.1, although there are more existing streamflow gauging stations with up-to-date records around the world than shown on this map. As pointed out by Hannah et al. (2011), the access to the measurements of other researchers or to official national-scale data sets is often limited due to various reasons, such as fear of misuse, national data policy, lack of time, awareness, knowledge or willingness to share. Therefore, many catchments are or can be viewed as either ungauged (Hrachowitz et al., 2013) or as no longer gauged (Hannah et al., 2011). Hence, further efforts are needed to achieve robust and reliable baseline data and predictions in developing countries (Hrachowitz et al., 2013), particularly where the need for water is greatest and in mountainous regions and the arctic were most freshwater sources are located (World Water Assessment Programme, 2003) but access is often difficult. Developing countries often rely on non-governmental organisations to build up measurement networks, but as gauging networks remain cost- and labour intensive, the money is often after a few years redirected to more pressing issues, such as disaster relief (Hannah et al., 2011).

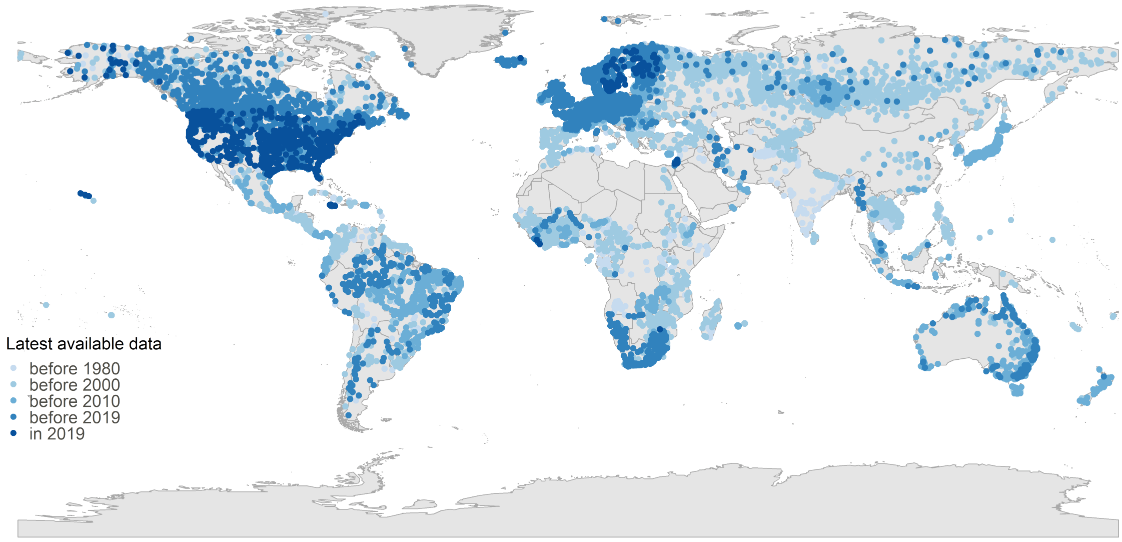



**Figure 1.1:** The number of runoff stations in the Global Runoff Database. The colour specifies the year of the last available measurement that was archived. The data were obtained in November 2019 from the Global Runoff Data Base: www.bafg.de and the base map was obtained from naturalearthdata.com using R.

There is a great demand for hydrological data that is freely accessible and simple to acquire, even in remote areas. Streamflow is still very hard to observe with a sufficient spatial and temporal resolution (Paper I) because gauging stations are expensive to construct and maintain. Existing alternative options include remote sensing (Smith et al., 1996), low-cost sensors (Peña et al., 2017), smartphone cameras (Le Coz et al., 2016) and webcams as suggested in van Meerveld et al. (2017). Citizen science has the potential to provide data in areas were no measurement infrastructure is available (Buytaert et al., 2014).

## 1.2 Citizen science

The Oxford English Dictionary[1] defines citizen science as

> *scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions*

and a citizen scientist as

> *a member of the general public who engages in scientific work, often in collaboration with or under the direction of professional scientists and scientific institutions; an amateur scientist*

Several attempts have been made to characterise the practices in citizen science projects and the degree of involvement of the citizen scientists in the projects: Bonney et al. (2009) divided projects into the three categories

- contributory: citizen scientists contribute data
- collaborative: the project is designed by scientists and citizen scientists help analyzing the data or are involved in the further design of the project
- co-created: citizen scientists and scientists work together, even in the project design phase

According to Strasser et al. (2018), this categorisation implies that projects that involve participants more in the design of the project, are to be preferred over projects that rely on citizen scientists for data collection only. To improve the categorisation scheme, Shirk et al. (2012) expanded the three categories by two other categoriese:

- contractual: researchers try to answer questions that were raised by the public
- collegial: contributions of e.g. amateur astronomers or birders who often make substantial contributions to their field

They stated that the five categories represent a spectrum where all categories are equivalent. Later, Haklay (2013) defined four levels of involvement:

- crowdsourcing: citizen scientists as sensors
- distributed intelligence: citizens as basic interpreters
- participatory science: participation in problem definition and data collection
- extreme: collaborative science - problem definition, data collection and analysis

[1] `www.oed.com` (accessed: 09.01.2020)

However, according to Strasser et al. (2018), these categorisations have a political agenda and aim to increase citizen empowerment. To avoid a political agenda, Strasser et al. (2018) proposed a new typology to characterise practices in citizen science:

- sensing (e.g. bird sightings)
- computing (e.g. by "donating" computing power)
- analyzing (e.g. online projects for image analysis or classification)
- self-reporting (e.g. of illness symptoms for medical studies)
- making (e.g. an open laboratory for citizen science)

A similar characterisation of the practices in citizen science was published in the *White Paper on Citizen Science in Europe* by Serrano Sanz et al. (2014), where the practices in citizen science are described as equivalent models of citizen engagement (examples my own, Figure 1.2):

- pooling of resources (e.g. by "donating" computing power)
- data collection (e.g. making water level class observations)
- analysis tasks (e.g. identification of species)
- serious games (e.g. gamified data collection)
- participatory experiments (e.g. citizens can conduct experiments with the help of scientists)
- grassroots activities (e.g. a research project started by citizens to assess the water quality in local households)
- collective intelligence (e.g. making use of the "wisdom of the crowd")

Serrano Sanz et al. (2014) do not explain the categories in further detail, but most categories overlap with those of Strasser et al. (2018). The categories *Serious Games* and *Collective Intelligence* refer to the use of the "wisdom of the crowd", such as in online games were multiple people classify the same image (e.g. Strobl et al. (2019)). This category might be implicitly included in the *analyzing* category of Strasser et al. (2018).

These different categorisation schemes all show that there are many ways that citizens and researchers can collaborate to solve problems that were defined by scientists or the public.

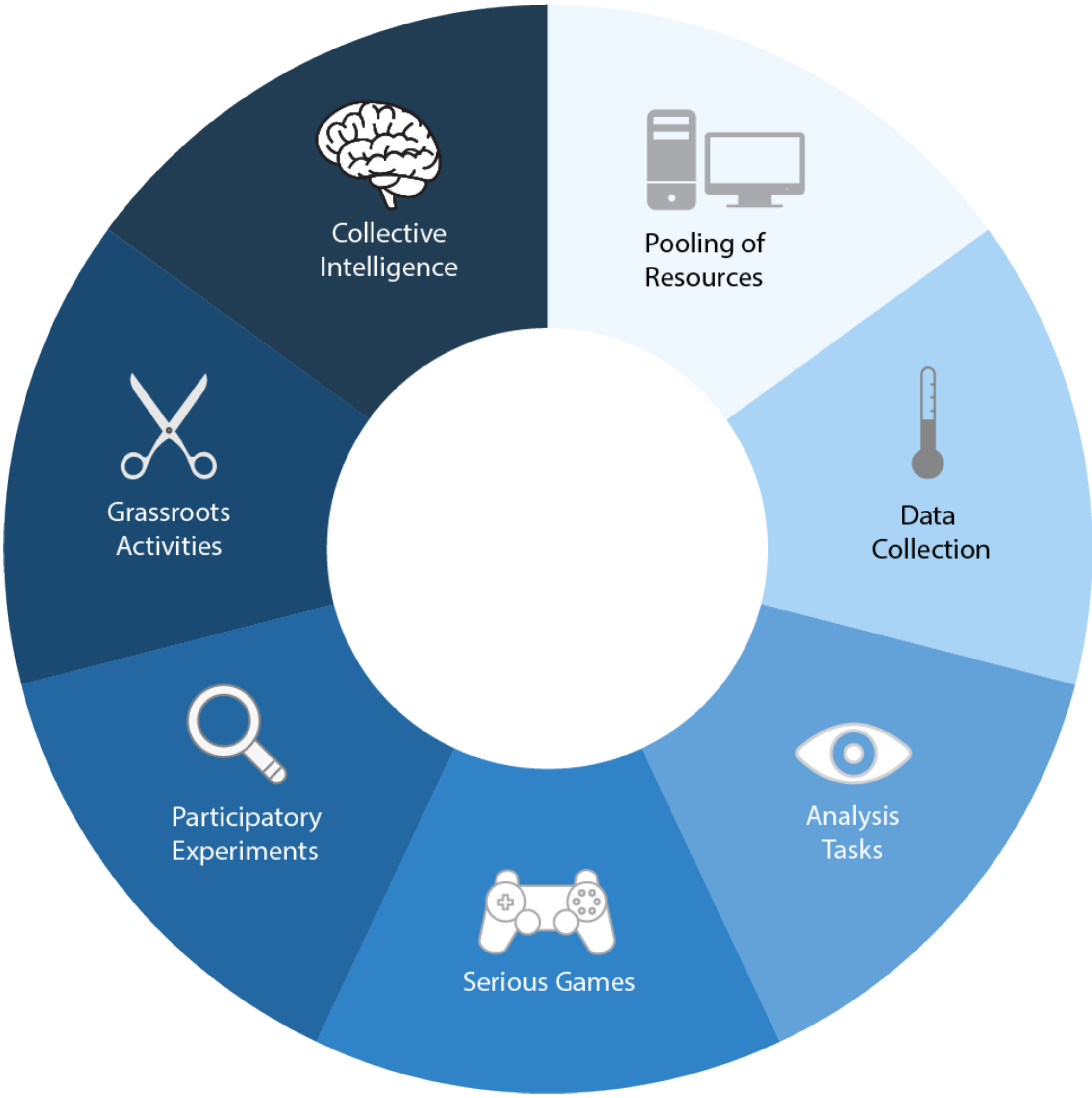



**Figure 1.2:** The spectrum of models of citizen engagement in citizen science projects in the *White Paper on Citizen Science in Europe* (Serrano Sanz et al., 2014). The different models are not explicitly explained in Serrano Sanz et al. (2014) but the graphic demonstrat the variability of engagement options that exist in citizen science projects. Adapted from Serrano Sanz et al. (2014).

### History of citizen science and existing projects

Involving citizens at different stages of research is not a new phenomenon. The Swedish meteorologist Tor Bergeron collected snow depth observations (Bergeron, 1949) and rainfall measurements using simple rain gauges (Bergeron, 1960). The data were sent by the citizens using postcards. One of the oldest (since 1900!) and still ongoing projects is the Audubon Christmas Bird Count, where every year around Christmas citizen scientists count birds (Meehan et al., 2019). However, there are even older examples of research that have characteristics of citizen science: in Japan, Aono & Omoto (1993) and Taguchi (1939) reconstructed the date of the cherry tree blossoming from old diaries and chronicles that date back to the 9$^{th}$ century.

Recent developments in smartphones and internet technologies, such as social media platforms offer new and exciting opportunities to include the public into research, for instance, by using crowd-based or volunteered geographic information (Capineri et al., 2016; Haklay, 2013), such as the analysis of tweets to determine the extent of earthquakes (Crooks et al., 2013). In hydrology, there are several flood related projects that rely on crowdsourcing or volunteered geographic information (See, 2019) from social media data, such as e.g. Twitter data (Arthur et al., 2018) or the PetaJakarta.org[2] project in Indonesia where people submit images and locations of floods and can at the same time ask for help (Ogie et al., 2019). Similar projects in Argentina, France and New Zealand ask citizens to send in videos and photographs from floods (Le Coz et al., 2016).

Other projects rely on more deliberate online participation of citizens. For example, GalaxyZoo[3] (Raddick et al., 2013) aims to identify shapes of galaxies by letting citizens compare a large number of images. The goal of the project Foldit[4] (Curtis, 2015) is to explore the numerous possibilities of protein folding.

There are also multiple outdoor projects that rely on the use of modern technology. For example the Austrian Naturkalender[5] project asks participants to collect phenological information, for instance, to document shifting start times of the blossoming of different plant species (Paper II). A very successful example of an environmental citizen science project is the Collaborative Community Rain, Hail, and Snow Network[6] in the United States (CoCoRaHS; Reges et al. (2016), where citizens buy simple rain gauges and report the rainfall amounts that they measure. Other examples of projects that involve more coordinated outdoor activities with citizen scientists include the collection of information on snow cover disappearance in the Pacific Northwest of the United States (Dickerson-Lange et al., 2016), the Great Pollinator Project[7], where citizens reported bee landings on designated plant species to assess the ecosystem quality for bees in New York City (Domroese & Johnson, 2017), and the project HydroCrowd were volunteers (mainly students) collected 280 water samples on a single day in Germany (Breuer et al., 2015).

[2] name changed to `https://petabencana.id/` (accessed: 21.04.2020)
[3] `www.galaxyzoo.org` (accessed: 09.01.2020)
[4] `https://fold.it` (accessed: 09.01.2020)
[5] `www.naturkalender.at` (accessed: 09.01.2020)
[6] `www.cocorahs.org` (accessed: 09.01.2020)
[7] `www.greatpollinatorproject.org` (accessed: 09.01.2020)

Other applications aim at monitoring of water quality in lakes, streams, rivers, wells, ponds, and wetlands (Conrad & Hilchey, 2011), e.g., by measuring water reflectance and turbidity with a smartphone camera in the HydroColor app[8] (Leeuw & Boss, 2018).

### Citizen science in hdrology

Since 2014 there has been an increase in citizen science-based studies in hydrology (Njue et al., 2019). Njue et al. (2019) report that out of all citizen science projects that aim at either the collection of water quality data, water level data, rainfall data or a mix of those, 63% are water quality related projects. Stepenuck & Genskow (2017) report that there are 345 such volunteer monitoring programs in the United States alone. Amongst the new projects there are also some water level and streamflow related projects. In CrowdHydrology[9], a project in the US (Lowry et al., 2019; Lowry & Fienen, 2013), passers-by read water levels from staff gauges in streams and submit them via text messages. Other projects with the same approach are Cithyd[10] in Italy and Weeser et al. (2018) in Kenya[11]. Weeser et al. (2018) showed that reimbursement for the costs of text messages increased participation rates and that the quality of water level observations read from physical staff gauges was reasonably good. Smartphones4Water[12] is a project in Nepal (Davids et al., 2017) that tested several simple streamflow measurement methods for citizens. Even though the approaches of the above projects worked quite well, they are not easily scalable as it is still costly and requires significant effort and time to install staff gauges or signposts at multiple sites. This PhD-thesis focuses on the CrowdWater project[13], which follows an approach that is similar to geo-caching and allows for an easier up-scaling of measurement in space to contribute to the collection of hydrological data in regions where such data are scarce.

### Motivation

The main motivations for people to join citizen science projects are to contribute to science and to protect the environment, as well as the feeling to belong to a community (Alender, 2016; Curtis, 2015; Raddick et al., 2013). The project's topic plays an important role as well because identification with the project's topic is important (Rey-Mazón et al., 2018; Frensley et al., 2017). Citizen scientists either want to learn something new (Domroese & Johnson, 2017) or help to solve issues that the project addresses (Johnson et al., 2014). These main motivations are often very similar in citizen science projects. However, there are only a few studies in Europe (e.g. Land-Zandstra et al., 2016) and none, that we are aware of, in Switzerland or Austria that address the motivations of citizen scientists. Therefore, Chapter 3 (and Paper II) investigates the motivations of

[8] `http://misclab.umeoce.maine.edu/research/HydroColor.php` (accessed: 09.01.2020)
[9] `www.crowdhydrology.com` (accessed: 09.01.2020)
[10] `www.cithyd.com` (accessed: 09.01.2020)
[11] `www.uni-giessen.de/hydro/hydrocrowd_kenya` (accessed: 09.01.2020)
[12] `www.smartphones4water.org` (accessed: 09.01.2020)
[13] `www.crowdwater.ch` (accessed: 09.01.2020)

CrowdWater participants and compares them to the motivations of participants of an Austrian citizen science project called Naturkalender.

## 1.3 The CrowdWater project

### 1.3.1 Introduction

The CrowdWater project started in 2016 and the smartphone application *CrowdWater / SPOTTERON* (hereafter referred to as the CrowdWater app) was launched in early 2017. The goal of the project is to develop a tool to collect hydrological information for hydrological models that can be used for flood warnings and other water management applications. Citizen scientists are asked to contribute pictures of streams and estimates of water level classes (WL-classes) based on a virtual staff gauge (Paper I, Seibert et al. 2019), to determine the state of temporary streams (Kampf et al., 2018), to estimate soil moisture based on qualitative classes (Rinderer et al., 2012), or to map plastic pollution in, and along streams in collaboration with *The Ocean Cleanup*[14]. Citizen scientists are encouraged to take repeated measurements at the same locations to obtain time series for these locations. The app functionalities for soil moisture, temporary streams and plastic pollution are described in other publications Kampf et al. (2018); Seibert et al. (2019); Rinderer et al. (2012). In this thesis, I focus on WL-class and streamflow estimates. Therefore the approach of the virtual staff gauge is explained here in more detail but see also Paper I.

### 1.3.2 Virtual staff gauge approach

The basic idea behind the approach to observe WL-classes is that it is usually possible to identify a number of features in a stream or on the stream bank, such as rocks, that allow ranking of the water levels (i.e., “below this tree but above that rock”). While such WL-class observations are not as precise as continuous water level observations from a staff gauge (i.e., no millimetre resolution) and provide more qualitative information such as “the water level is very low” or “there is a flood event,” they can be quite informative for hydrological modelling (van Meerveld et al., 2017). The challenge is to allow a simple identification of the different WL-classes, without the need for lengthy verbal descriptions. A picture is helpful in this respect but needs to be amended by a scale. For this, we use the virtual staff gauge approach (Figure 1.3). In practice, this means that the citizen scientist takes the following steps:

- The user chooses a suitable location along a stream and identifies it on a map in the smartphone app.
- The user takes a picture of the streambank (perpendicular to the flow direction and as level as possible, to minimize contortion of the view). There should be some

[14] `www.theoceancleanup.com` (accessed: 09.01.2020)

reference in the picture, such as a bridge or stones and ideally, the picture is taken during low flow conditions.

- An image of a yardstick with ten classes is digitally inserted into the picture as a virtual staff gauge. The user can move this virtual staff gauge in the image and scale it so that it is level with the current water level and covers the expected stream level variations.


**Figure 1.3:** An example reference image with the virtual staff gauge inserted in it, taken in Chosica, a village located upstream and east of Lima in Peru. Photo taken by Renato Gazzola. CrowdWater Spot: spotteron.com/crowdwater/spots/21414 (accessed 09.01.2020).

This reference picture with the virtual staff gauge allows anyone who visits the site at a later time to estimate the WL-class by comparing the current water level to the features on the photo and the virtual staff gauge (e.g., the water level has changed and is now above a certain rock). More specifically, the user compares the current water level with the reference picture with the staff gauge in the app, takes a new picture of the stream, selects the current WL-class on the horizontal staff gauge (Figure 1.4) and submits the new observation to the data server. For details on the design of the virtual staff gauge the reader is referred to Paper I.

When repeated observations are submitted for the same location, this results in a time series of water level class observations. It is important to note that the user observes and enters the WL-class; the new picture is only used for documentation. While automated image recognition could be valuable, at this point we rely on human eyes and interpretation to avoid issues related to the exact location and angle when the picture is taken. The pictures, however, allow data quality control. We have developed the CrowdWater game as an approach to use these pictures for crowd-based quality control of the WL-class data (Strobl et al., 2019).

Typical errors in placing the virtual staff gauge are related to the size of the virtual staff gauge, its placement, and the angle of the photograph. These mistakes affect about 10% of the more than 500 reference pictures that were made by the time Paper I was written. Staff gauge placement or size problems could be due to users not having read the available instruction material or not fully understanding the concept. Other issues are not directly related to setting up a virtual staff gauge site but still affect the results, e.g., it is less useful if users create new measurement sites in, or close to, a location where another spot already exists than when they update the existing spot or start a new site on a different river.

### 1.3.3 Number of participating citizen scientists and contributions

The smartphone application is designed to become a social network, where users can follow each other, like, comment and share contributions. These functions have however, so far not been used widely by the participants. Most of the social interaction in the CrowdWater app occurs between the project team and citizen scientists via the comments function or by personal communication via e-mail. Only in rare cases do citizen scientists comment on other observations. The CrowdWater project has so far mainly been advertised via social media (Facebook, Twitter, and Instagram) and in our private and work-related networks (e.g., presentations at conferences, schools and science fairs, articles in university newsletters and magazines, a press release by the University of Zurich etc.). Most of the advertisement and outreach for the CrowdWater project focused on German speaking citizens, hence most data have been collected in Switzerland and Austria. However, observations can – and have been made – around the globe. Since the value of the data is still subject to research, communication regarding the potential use of the data for flood warning systems has been done rather carefully. By the time of writing this thesis in January 2020 there were 580+ participants who contributed at least one observation for one of 2'700+ unique spots. In total, there were 10'900+ contributions (Figure 1.5) of which 5'200+ were water level observations, 900+ were soil moisture observations, 4400+ were intermittent stream observations, and 400+ were observations for plastic pollution.

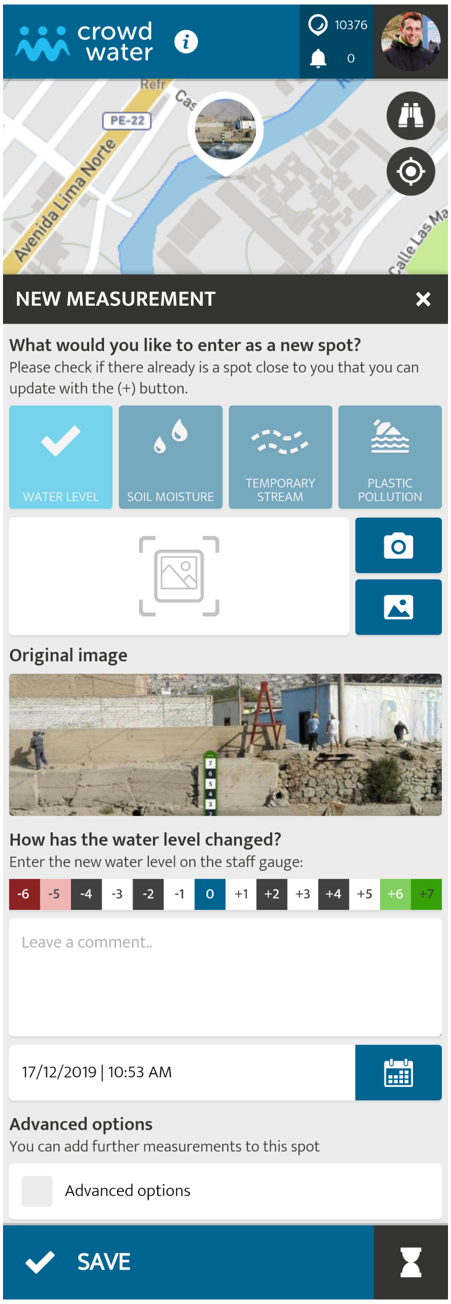



**Figure 1.4:** Screenshot of the CrowdWater app of the screen for entering a new water level class observation. The observed water level class can be entered by clicking on the number in the horizontal staff gauge. Uploading a new photo is optional but encouraged. Streamflow estimates can be made when the *Advanced options* are selected.

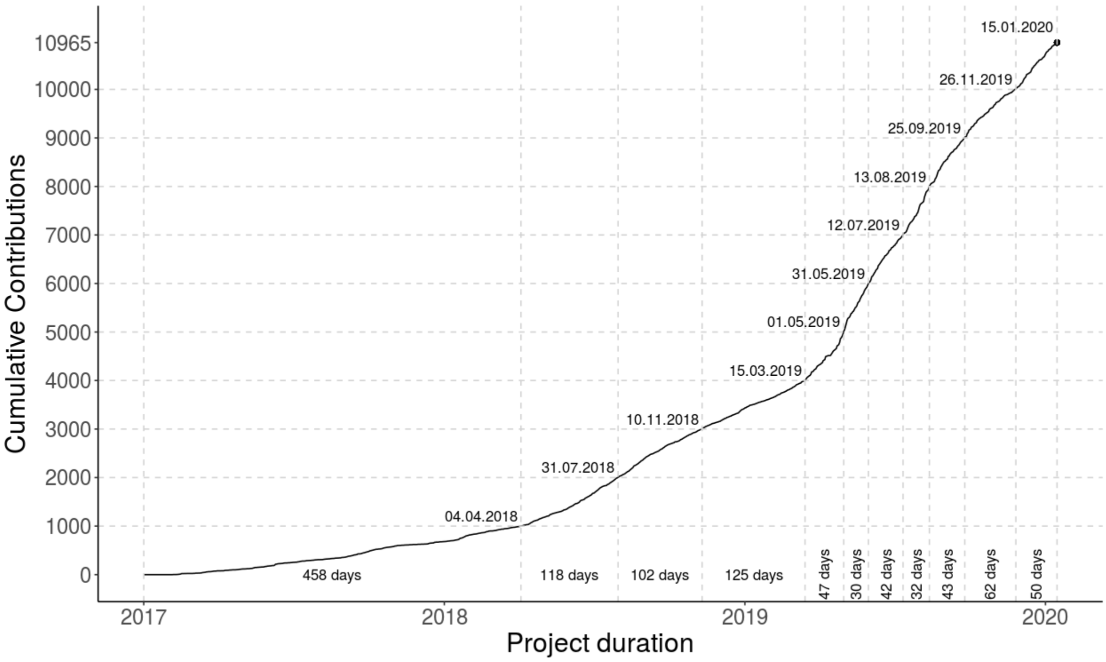



**Figure 1.5:** Cumulative number of contributions to the CrowdWater project. Figure generated by the CrowdWater dashboard on crowdwater.ch/dashboard. Accessed: 15.01.2020

# 2

# Scope of the thesis and research questions

## 2.1 Scope of the thesis and research questions

The approach to use citizen science for the collection of hydrological data is not new. However, the approach that is used in the CrowdWater project, particularly the use of crowd-based estimates of streamflow and WL-class estimates with virtual staff gauges in the reference images (Paper I) has not been evaluated before. This thesis focuses on this part of the CrowdWater project, although I also participated in the development of the measurements for the other variables in the app. This thesis mainly contributes to the knowledge on the motivation of participants in environmental citizen science projects by asking respondents what motivated them initially to join CrowdWater and Naturkalender and how these initial motivations were fulfilled by their participation. The hydrological side of the thesis contributes to the knowledge on the accuracy and the value of hydrological data that are collected in a simple manner by comparing the estimates to measured data and by investigating the information content of crowd-based streamflow and WL-class estimates for hydrological model calibration. More specifically the thesis addresses the following research questions:

1. **What motivates the participants of CrowdWater and Naturkalender to join these projects and in how far are these motivations fulfilled by participation?** In an online questionnaire, I asked participants of the citizen science projects CrowdWater and Naturkalender what had motivated them initially to join these projects and which of these motivations had been fulfilled by their participation. The data were evaluated based on two different frameworks on motivation in citizen science and volunteering from the literature and are described in Chapter 3 and Paper II.

2. **How good are streamflow and water level class estimates by citizen scientists?** This question was addressed in two different studies: a survey at the start of the CrowdWater project when the number of actual data submissions was very small and an analysis of the data collected using the CrowdWater app and forms near multiple streams. In 16 field surveys at the start of the project, we tested three simple methods to estimate water quantities in streams that could potentially be used in a citizen science project and do not require any equipment. We asked passers-by at 10 streams in the greater area of Zurich in Switzerland to estimate the streamflow directly or via width, average depth and flow velocity. Additionally, we asked them to estimate the water level class by comparing the current water level to a reference image with a virtual staff gauge. We then compared the estimates of citizens to measured streamflow quantities to compare their accuracy. This is work is described in Chapter 4.2 and Paper III. In the second study, we compared the crowd-based time series of water level classes with water levels that were measured in the vicinity. We did this for nine measurement locations where data were collected with the CrowdWater app and twelve field stations where people could report the water level class on paper forms. This work is described in Chapter 4.3 and in Paper IV.

3. **What is the potential value of crowd-based streamflow and WL-class time series for hydrological model calibration?** The error distributions of the streamflow and water level class estimates from the surveys (Paper III) were used to create synthetic streamflow (Paper V) and water level class (Paper VI) datasets that have uncertainties that are typical for citizen science data. I sub-sampled these datasets to create synthetic datasets with different temporal resolutions that represent scenarios with different contribution times and frequencies ranging from hourly estimates to one estimate per month. We calibrated the hydrological model HBV-light with these datasets for six (Paper V) and four (Paper VI) catchments in Switzerland and evaluated the performance of the model for different years by comparing the simulated streamflow with the observed streamflow. The citizen science-like streamflow and water level class observations were considered to be valuable for hydrological model calibration if the validation performance for the model calibrated with this data was better than that of model runs with random parameter sets. The results for these studies are described in Chapter 5 (and Paper V and Paper VI).

# 3

# What motivates citizen scientists to contribute to the CrowdWater and Naturkalender projects?

## 3.1 Introduction

It is important to understand the different motivations of participants in citizen science in order to attract participants and to lower the hurdles for sustained participation. The motivations that drive people to participate in citizen science and what people gain from participation are, however, complex (Strasser & Haklay, 2018; Thornhill et al., 2019; West & Pateman, 2016). The main motivations to join citizen science projects, reported so far, are to contribute to science and to protect the environment, as well as the community aspect (Alender, 2016; Curtis, 2015; Raddick et al., 2013). However, many studies so far focused on a single project and used only one classification scheme to analyse the results. The use of different approaches and surveys to assess the motivations of participants, the different schemes to classify the motivations with different levels of detail, and the substantial differences in the projects make it difficult to compare the results of the different studies on motivations to participate in citizen science projects. We aimed to expand the knowledge on the motivation of citizen scientists by comparing the motivations of participants in two projects: CrowdWater and Naturkalender (English: Nature's Calendar). Naturkalender is a smartphone based, environmental project based in Austria and aims to document the phenology of indicator plant species and the occurrence of indicator animal species to detect potential changes in response to climate change. The two projects have, so far, mainly recruited participants from western European countries (most of the participants come from Switzerland and Austria). The comparison of the motivations to participate in the two projects enables a more explicit focus on how the project topic, thematic content and outreach activities affect the motivations of the participants because the projects are similar in terms of the visual design of the app, the way data are transmitted, and cultural background of the participants. The full description of the study can be found in Paper II and a simplified graphical summary is given in Figure 3.1.

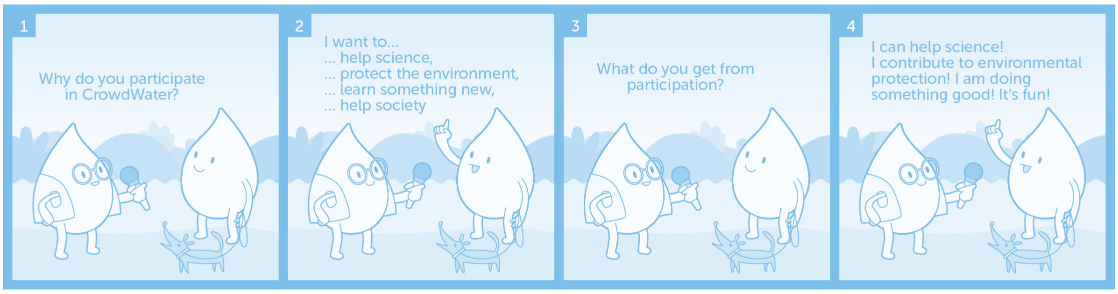
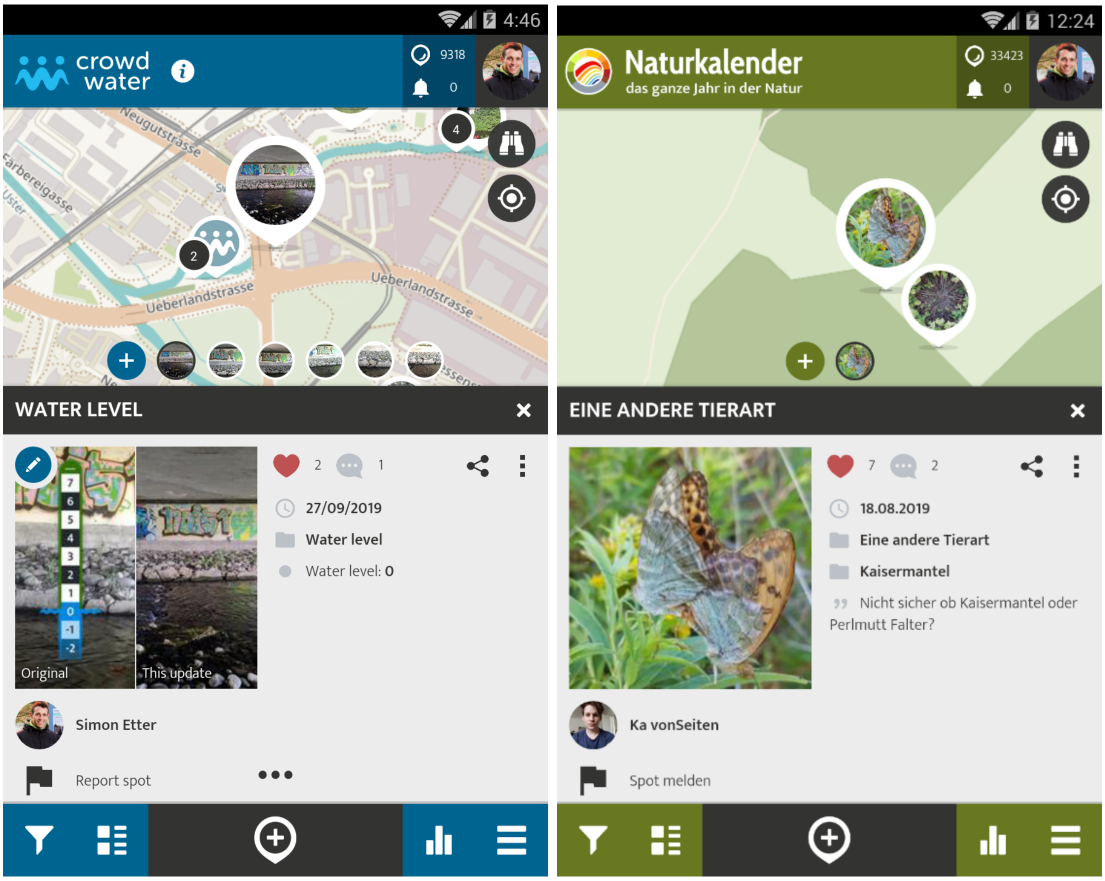



**Figure 3.1:** A simplified illustration of the questions posed and the answers given in the online questionnaire on motivation in Paper II. Design by: University of Zurich, Information Technology, MELS/SIVIC, Tara von Grebel

## 3.2 Methods

To assess the motivations of participants in the projects, we invited about 400 people, who had registered for the CrowdWater newsletter by e-mail, and additionally used the push-message service in both apps to fill out the questionnaire. The questionnaire contained 29 statements that were based on the scientific literature on motivation in citizen science (Levontin et al., 2018). The statements give potential reasons for why people joined a citizen science project. We used these statements in the first part of the questionnaire to ask what motivated people to join the projects – the engagement part. For the second part of the questionnaire, the fulfilment part, we rephrased most of the statements to ask whether these motivations were fulfilled by the participation in the project. Answers were given on a Likert scale with five options that were translated in numbers: *don't agree at all* = 1, *slightly disagree* = 2, *undecided* = 3, *slightly agree* = 4, *fully agree* = 5. We received answers that we could use for the study from 54 CrowdWater participants and 36 Naturkalender participants. We classified the statements according to the scheme of Batson et al. (2002), which was adapted by Beza et al. (2017) and is hereafter referred to as the *Batson-scheme*, to obtain an overview of the broad categories of motivation. Additionally, we used the scheme of Schwartz et al. (2012), which was adapted for citizen science projects and recently published in a questionnaire by Levontin et al. (2018), hereafter referred to as *Schwartz-scheme*, to gain more detailed insights for the entire spectrum of motivations (see Figure 2 in Paper II). We chose these two frameworks because they cover the broadest range of potential motivations. A full list of the statements and the categories of the two frameworks can be found in Paper II. We used the paired Wilcoxon signed rank test to test the significance of the differences between the median response to the statements regarding the motivations for initial




(a) CrowdWater App (b) Naturkalender App

**Figure 3.2:** Screenshots of the CrowdWater (a) and the Naturkalender app (b), with on the top row of the second panel of each screenshot the social media features (from left to right the like button and counter, the speech bubble that allows users to comment on the observation (with the counter next to it), and the sharing button to share contributions on Facebook, Twitter and Google+. More information on the app design can be found in Paper I, Seibert et al. (2019) and spotteron.net (accessed: 09.01.2020). Figure from Paper II.

.

engagement and the fulfilment of these motivations by participating in the projects. We used the Mann-Whitney U-test to test the significance of the differences in the median response for the different subgroups of respondents (i.e. CrowdWater vs. Naturkalender participants, super-users who contribute on average at least once per week vs. occasional participants, and the different age groups).

## 3.3 Results

Participants of the CrowdWater and Naturkalender projects mainly joined the projects to contribute to science, satisfy their interest in science and technology, protect nature, contribute to the well-being of society, learn something new, and be physically active (Figure 3.3).

Not all the initial motivations were fulfilled by participating in the projects (Figure 3.4). The respondents of both projects, for instance, agreed significantly less that their continued involvement was driven by a motivation to contribute to society (*universalism, societal concern*) and socialising with other people (*security and belongingness*) although these aspects motivated them to join the projects. On the other hand, fun and enjoyment (*hedonism*) were not the primary motivation to become involved in the projects but were essential motivators for continued participation. Respondents from Naturkalender were more motivated by enjoyment, learning (*self-direction*) and being outdoors and the physical activity (*stimulation*) than the CrowdWater respondents. Most of the fun and learning experience probably came from the social interaction and the information on plants and animals included in the Naturkalender app. Such a learning aspect was not available for CrowdWater, which probably explains why for CrowdWater respondents the primary motivation for continued participation was similar to the engagement motivations: help with research (*universalism, research*), protection of nature (*universalism, nature*) and acting according to their values and beliefs (*tradition*).

## 3.4 Conclusions and implications

From a combination of the findings in Paper II and the literature, we could draw the following conclusions and recommendations for involving citizen scientists in research projects:

- It appears that people are more likely to contribute to a project over extended time periods if they have shared values with the project's goal (e.g. protection of the environment). The level of interest increases if projects tackle problems that impact the every-day life of participants (Frensley et al., 2017). One could argue that everyone, and thus also Naturkalender participants, is affected by climate change and people can observe the effects in their backyard. For CrowdWater, the local relevance of the stream observations is less evident because the data are not linked to any forecasts (yet). The motivation to participate in CrowdWater might change, once the project is more frequently used in other countries with fewer

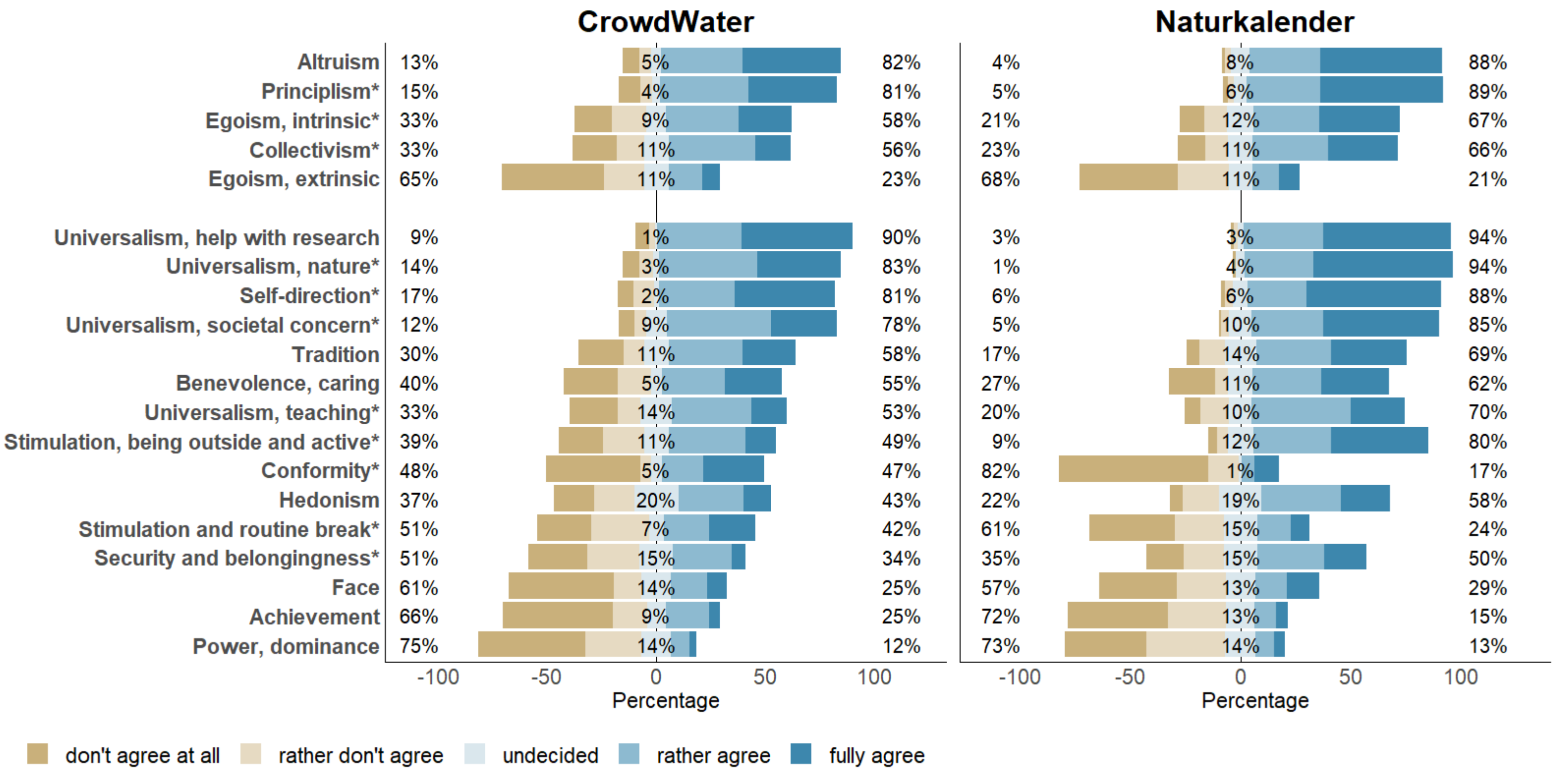



**Figure 3.3:** Percentage of respondents who chose one of the five levels of agreement to statements regarding initial engagement that belong to the motivational categories of Batson et al. (2002) (top five rows) and Schwartz et al. (2012) for CrowdWater (left) and Naturkalender (right). For the categories marked with an asterisk (*) the median response for the CrowdWater and Naturkalender participants was significantly different. The values next to the categories indicate the percentage of respondents who don't agree (left; don't agree at all and rather don't agree), are undecided (middle) and agree (right; rather agree and fully agree). The categories are sorted by decreasing percentage of agreement for the respondents of the CrowdWater project. Figure from Paper II.

gauging stations and/or locations where people are more exposed to water-related hazards.

- Participants need to be interested in the topic of the project and the activities involved. They often have an interest in science or technology. For online projects, the motivation to participate in a project is mainly to contribute to science (Curtis, 2015; Raddick et al., 2013). For Naturkalender, it seems that many participants are plant (and animal) enthusiasts. The agreement to the statement "I am interested in the topic of this project" was very high for respondents for both projects, similar to the findings of Hobbs & White (2012) for two wildlife observation projects.

- Social media elements are beneficial for online projects (Nov et al., 2014) to create social networks and allow people to comment on the contributions of others. This could help to form a self-organizing community that ensures data quality (Serret et al., 2019). This is in line with self-determination theory, according to which the ability to make competent actions and decisions autonomously leads to enhanced self-motivation (Ryan & Deci, 2000). In Naturkalender, social interactions enable participants to help others and therefore provide teaching and learning experiences for the participants without requiring effort by the project administrators. In CrowdWater this feature is not used extensively, perhaps this is related to the lack of options to share knowledge and learn something new.

- The importance of learning new things has been reported in multiple studies (e.g. Hobbs & White, 2012; Johnson et al., 2014). For Naturkalender, *self-direction* was the category with the second highest agreement (average: 86% agreement) in the fulfilment part, whereas for CrowdWater it was only ranked 6$^{\text{th}}$ (66% agreement). CrowdWater offers information about hydrology on the homepage and links to an online course called “Water in Switzerland”. However, so far it appears that these options are rarely used, possibly due to them being mentioned on the homepage, rather than within the app. Thus, opportunities for learning are limited compared to Naturkalender, where users profit from the expertise of other participants and informative content on plant and animal species inside the app. For successful projects, there should be an easily accessible possibility to extend one’s knowledge about a topic and to learn new things.

- People need to enjoy their participation. This can be achieved by providing more choices and options for participating, as it is the case in Naturkalender compared to CrowdWater, but also by giving users more competences (e.g. more rights for advanced users) as proposed in the self-determination theory (Ryan & Deci, 2000). The option to give selected users the right to edit contributions of other users exists in the CrowdWater app, but has not been used so far.

- The super-users were in general older than the occasional participants. This is common for other projects as well (Sheppard et al., 2017; Wright et al., 2015). It might therefore be an effective strategy to focus recruitment on people above the age of 50. This was to some extent unexpected because both projects use modern apps, which might be less intuitive for some older people. On the other hand, once the habit is established, older people are more likely to contribute for extended periods (Sheppard et al., 2017; Venkatesh et al., 2012).

- Participants of the newer CrowdWater project were considerably more motivated to join by social pressure (*conformity*), i.e., because they were asked to help with the project. This might be true for many projects that have just started and still rely on families, friends or acquaintances to participate in (and promote) the project. People who were motivated to join by a perceived social pressure may help a project in the beginning but later tend to quit. Naturkalender participants were motivated more to join because of their interest in the project topic, in combination with a willingness to share their expertise on the topic.

- The introduction of gamification elements increases the competitive element (Nov et al., 2014) and projects might reach new audiences (Bowser et al., 2013b) but this might also decrease the intrinsic motivation of participants (Thiel & Fröhlich, 2017) or cause participants to make low-quality contributions in order to get more points (Bowser et al., 2013a). Thus, gamification should be applied cautiously and potential negative consequences should be evaluated beforehand. The respondents of this survey agreed relatively little with competitive categories (achievement, face). Whether people did not like the existing leader board, or if it was not enough to trigger these motivations, remains to be investigated.

**Implications for CrowdWater**

For the future of CrowdWater, it is important to recruit participants that contribute over longer periods and can thus collect long time series. Therefore, groups of people who are interested in, or affected by water need to be identified. Such groups could then be moderated by a member of the same group, by e.g. giving that person more rights in the app. Adding more options to contribute easily and frequently might lower the barriers for contribution. Furthermore, adding informative content to the app that contributes to the learning experience of the participants could recruit or retain participants that are motivated by self-direction. The learning experience could be further intensified by providing more feedback on contributions from the project administration, as well as from other users in the comment sections of the app. Sharing research results would probably help the participants relate their contributions to research and thereby fulfil expectations related to helping research and in turn help to enhance participation and retention rates.

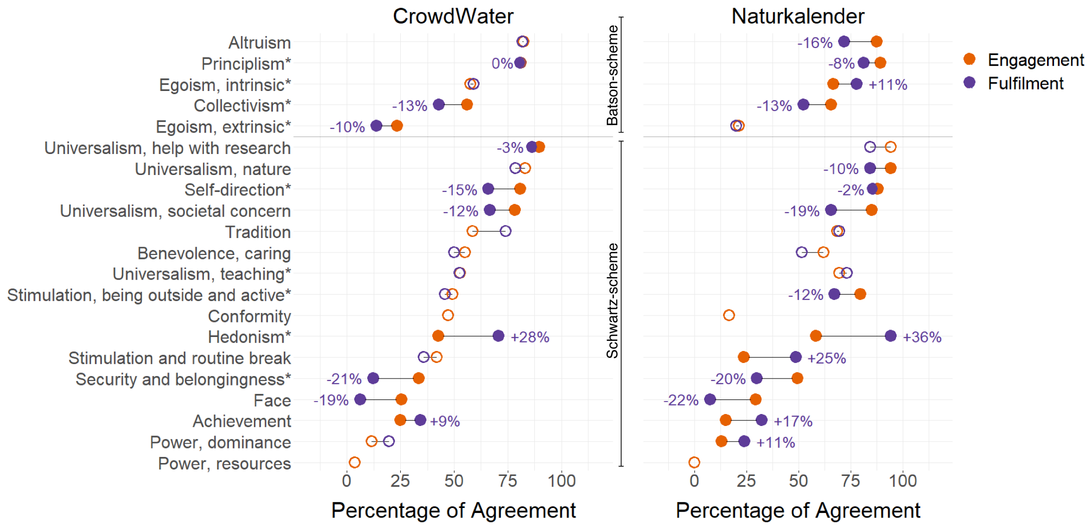



**Figure 3.4:** The average percentage of respondents that agreed to the statements that belong to the different categories for the motivations for initial engagement (orange) and fulfilment (purple) for CrowdWater (left) and Naturkalender (right). Empty circles indicate insignificant ($p>0.05$) changes in the median response for initial engagement and fulfilment; filled symbols indicate significant changes. Asterisks indicate categories for which the median response for fulfilment for the CrowdWater and Naturkalender participants was significantly different (see Figure 3.3 for the statically significant differences in agreement for initial engagement). The categories are sorted by decreasing agreement for the CrowdWater respondents in the engagement part. Figure from Paper II.

# 4

# What is the accuracy of crowd-based streamflow and water level class estimates?

## 4.1 Introduction

Determination of the accuracy of citizen science data before starting a citizen science project ensures that the data collected are sufficiently accurate for the purpose of the project. It furthermore avoids unnecessarily burdening citizens with tasks that result in data that are in hindsight of limited value due to data accuracy issues. We therefore conducted 16 field surveys at the start of the CrowdWater project in 2016 and 2017 (i.e. before the smartphone app was released in Spring 2017) to determine what types of parameters related to streamflow citizens can estimate accurately and to assess if streamflow or WL-classes are estimated more accurately. The full description of the study can be found in Paper III and a simplified graphical summary is given in Figure 4.1.

We conducted a second study on the accuracy of WL-class estimates that were collected with the CrowdWater app between spring 2017 and fall 2019 and with paper forms between fall 2016 and fall 2019. These crowd-based WL-class estimates were contributed by independent and non-supervised citizen scientists. Thus, this study extends the findings of Paper III where the experts were present. We selected nine locations where citizen scientists independently contributed data with the CrowdWater app and twelve locations where they could use paper forms and letter boxes. For all these locations, which were distributed across Switzerland and Austria, ‘true’ water level measurements from official agencies, ourselves or other research groups were available in the vicinity. Furthermore, we studied the temporal patterns of the contribution by citizens and discuss experiences from the two different approaches (app and forms) for data collection used within the

CrowdWater project. The full description of the study can be found in Paper IV and a short summary is given in Figure 4.2.

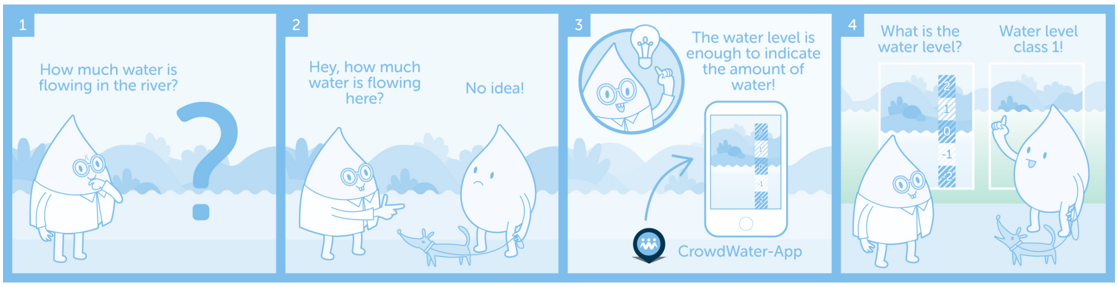



**Figure 4.1:** A graphical summary of Paper III depicting the street surveys with passers-by. The comic illustrates that water level classes are easier to estimate than streamflow for citizen scientists. Design by: University of Zurich, Information Technology, MELS/SIVIC, Tara von Grebel

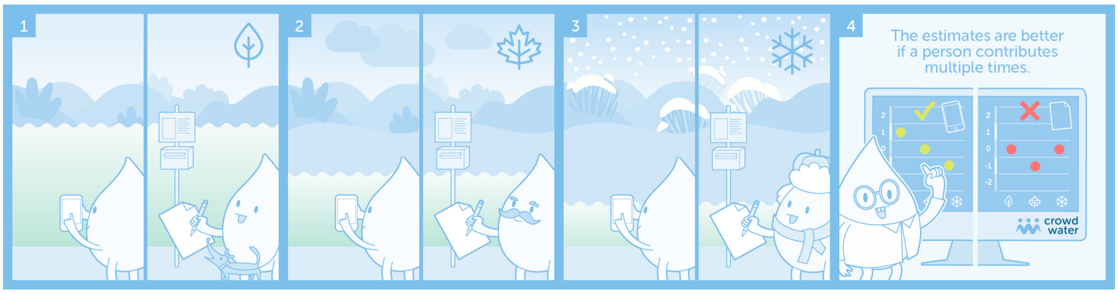



**Figure 4.2:** A graphical summary of Paper IV showing the different seasons when citizen scientists contributed to the app spots and the pen-and-paper stations and the difference in data quality between the two approaches. Design by: University of Zurich, Information Technology, MELS/SIVIC, Tara von Grebel

## 4.2 Field surveys

### 4.2.1 Methods

The aim of the surveys was to obtain a sufficient number of streamflow estimates for a specific stream on a specific day (our aim was 30 participants per survey to assure statistical significance; Field et al., 2013). The CrowdWater project aims to collect observations for the same stream at multiple times, but here we collected multiple estimates at (almost) the same time for the assessment of the accuracy of the estimates and we assumed that the streamflow remained constant during the survey. We thereby could assess the accuracy of the estimates compared to a nearby measurement for the same stream which was assumed to be correct. We conducted 16 field surveys where we asked 517 citizens to estimate the streamflow, as well as the average width, depth and velocity of the stream, and the WL-class. For the surveys, we selected 10 locations (Table 4.1, for more details see Table 1 and Fig. S1 of Paper III), where we expected enough people to pass by and to have time for the survey. We used a logistically simple sampling strategy, whereby we personally approached passers-by (similar to Breuer et al. 2015) and asked if they would complete the 5-minute survey (i.e., we did not use a targeted approach to capture responses of a representative group of citizens). No data were collected on the percentage of passers-by who participated, but we estimate that about every third person we approached agreed to participate in the survey. In addition, we asked high-school (Magliasina) and university students (Chriesbach, Glatt and Limmat) to fill out the survey during excursions. All surveys took place between October 2016 and September 2017. In total, we received 517 complete surveys: 372 passers-by, 61 participants from a geography bachelor student excursion (Glatt and Chriesbach), 40 from a high-school student excursion (Magliasina) and 44 from a summer school for PhD students from fields ranging from physics to social sciences (Limmat; Table 1). During the group excursions, we emphasized the need for individual estimates and limited discussions between the students for the duration of the survey. Participants were first asked to estimate the streamflow directly. For this direct estimate, we asked them to estimate the flow in $m^3/s$, or in $L/s$ for the very small streams. After this initial guess of the streamflow, we explained to the participants that it is possible to estimate the individual factors (width, mean depth and flow velocity) and to derive the streamflow by multiplying these values. The participants were then asked to estimate the average width, mean depth and velocity of the stream. We also asked participants to estimate the WL-class: the participants compared the current water level with a printed photo of the same stream (taken at an earlier time) with the virtual staff gauge with 10 WL-classes as it is used in the CrowdWater app (Paper I). We converted the WL-classes into streamflow ranges to make the accuracy of WL-class estimates comparable to the accuracy of streamflow estimates: For the stream locations with a nearby gauging station of the Swiss Federal Office for the Environment (FOEN; Sihl, Limmat, Aare), the classes of the virtual staff gauge were converted to a metric value by determining the stream depth that corresponded to each WL-class (i.e., mid-point and upper and lower water level for each class). We used the FOEN rating curve to convert these water levels to a streamflow estimate. For the sites where no rating curve was available (Hornbach, Irchel, Schanzengraben and Töss), additional measurements of the stream profile and water surface slope (estimated based on the slope of the streambed) were used to estimate the streamflow for each WL-class using the Manning-Strickler formula (Manning, 1891). This curve was fitted to the streamflow measured on the day of the surveys by adjusting the roughness coefficient within predefined boundaries based on the streambed material. Since the WL- classes represent a range of values rather than just one value, the streamflow was not only calculated for the centre value of the class, but also the class boundaries to obtain the possible range of streamflow values. The estimates from Chriesbach, Glatt and Magliasina were excluded from this analysis (101 of the 517 estimates) because the relevant data were not collected at the time of the surveys.


**Figure 4.3:** Screenshot of the CrowdWater app at the Salzach in Salzburg, where most contributions were made by the one user (Karin Ebermann). The image labelled *Original* shows the virtual staff gauge and the image labelled *This update* shows a later contribution with the same WL-class. Right: A pen-and-paper station at the official gauging station Kleine Emme – Werthenstein of the Swiss Federal Office for the Environment (in the background of the image) where 45 different citizens contributed observations. Note the reference image with the virtual staff gauge on the lower left of the sign.

### 4.2.2 Results

As expected, estimation of the individual streamflow factors with width, mean depth and flow velocity led to more accurate streamflow estimates than the direct estimates of streamflow (Figure 4 in Paper III). The main reason for this difference was the unfamiliarity of the participants with the units of $m^3/s$ for the direct estimates. However, there was still a large spread in the streamflow estimates based on the individual factors, as especially the depth was hard to estimate. In Figure 4.4(a) the spread in the estimated streamflow is shown for the medium sized rivers. These rivers are selected here, because the resulting error distribution from this study was used for Paper V.

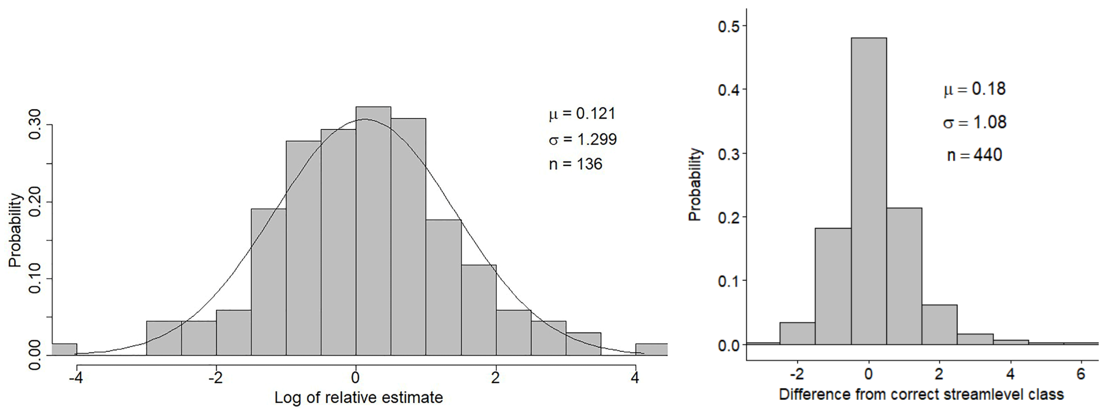
The WL-classes were estimated correctly by about half of the participants (48%) and most of the remaining participants (40%) were off by only one class. There were only a few outliers: 13% of participants had an error of two classes or more (Figure 4.4(b)). The largest overestimation was six classes and the largest underestimation was three classes. These errors likely occurred due to a misunderstanding of the method. The WL-class estimates were especially accurate for smaller streams where the streambank on the opposite side of the stream, where the virtual staff gauges were located in the photo, were close to the participant (Figure 5 in Paper III). One of the very small streams (Irchel) had a poorly placed staff gauge. The image was taken looking down onto the

**Table 4.1:** (adapted) Information on the streams where the field surveys took place. Size classes XS: $\leq 1$ m$^3$/s; S: >1–50 m$^3$/s, M: >50–200 m$^3$/s and L: >200 m$^3$/s. Survey dates are given as dd.mm.yyyy. A map with the survey locations is given in the supplementary material of Paper II (Fig. S1).

| Stream | Size | No. of survey | Date | n participants | Streamflow [$m^3/s$] | Approx. distance to virtual staff gauge [m] | Comments |
|---|---|---|---|---|---|---|---|
| **Chriesbach (Zurich)** | XS | | 29.09.2017 | 30 | 0.38[1] | 5 | BSc students: no direct streamflow estimates |
| **Hornbach (Zurich)** | XS | | 19.02.2017 | 33 | 0.134[1] | 8 | |
| **Irchel (Zurich)** | XS | | 11.03.2017 | 25 | 0.01[1] | 1 | |
| **Glatt (Zurich)** | S | | 29.09.2017 | 31 | 2.8[2] | 11 | BSc students: no direct streamflow estimates |
| **Magliasina (Magliaso)** | S | | 28.04.2017 | 40 | 16[3] | 14 | High school students: no stream level class estimates |
| **Schanzengraben (Zurich)** | S | | 01.04.2017 | 31 | 2.6[1] | 16 | |
| **Sihl (Zurich)** | S | 1 | 18.02.2017 | 33 | 7[3] | 32 | Low flow |
| | | 2 | 26.07.2017 | 31 | 28[3] | | High Flow |
| **Töss (Winterthur)** | S | | 12.03.2017 | 35 | 9[2] | 29 | Interpolation between three nearby stations for reference value |
| **Limmat (Zurich)** | M | 1 | 29.10.2016 | 38 | 59[3] | 7 | No streamlevel class estimates |
| | | 2 | 08.04.2017 | 27 | 83[3] | | |
| | | 3 | 02.06.2017 | 31 | 107[3] | | |
| | | 4 | 09.07.2017 | 44 | 75[3] | | PhD students, low flow |
| | | 5 | 13.11.2017 | 31 | 222[3] | | High flow |
| **Aare (Brugg)** | L | 1 | 07.01.2017 | 27 | 108[3] | 53 | Low flow |
| | | 2 | 10.05.2017 | 30 | 389[3] | | High flow |

[1]Streamflow data were obtained using salt dilution gauging.
[2]Streamflow data were obtained from the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA; hydrometrie.zh.ch)
[3]Streamflow data were obtained from the Swiss Federal Office of the Environment (FOEN; hydrodaten.admin.ch)

stream rather than horizontally from the height of the water level, which distorted the virtual staff gauge relative to the wall behind the stream, which made it more difficult to read. The accuracy of the WL-class estimates was better for the Limmat than for the Aare, even though they have similar widths (50 and 52m) and were the widest streams in the study. At the Limmat the virtual staff gauge was placed on a bridge pillar which was relatively close to the observer, whereas at the Aare it was placed on the opposite bank. From this we conclude that the virtual staff gauge, or rather the reference structures which are needed to select the WL-class should be close to the observer and that the placement of the virtual staff gauge is important.




(a) (b)

**Figure 4.4:** (a) Fit of the normal distribution to the frequency distribution of the log transformed relative streamflow estimates (ratio of the estimated streamflow and the measured streamflow) for the medium sized streams. This error distribution was used in Paper V. Figure from Paper V.
(b) Distribution of the errors in the WL-class estimates (i.e., the difference between the reported WL-class and the actual WL-class, as determined by experts) from field surveys for nine different locations. This error distribution was used in Paper VI. The virtual staff gauge used in the survey had ten classes. Figure from Paper VI

.

## 4.3 Real CrowdWater data

### 4.3.1 Methods

We selected nine locations in Austria and in Switzerland where multiple crowd-based WL-class estimates from the CrowdWater app were made (hereafter referred to as *spots*) and measured water level data were available for more than one year (Figure 2 of Paper IV). The spots had between 46 and 505 contributions at the time of this study in October 2019. We also installed signs with reference images (Figure 4.3) at twelve different stream locations in Switzerland (Figure 3 of Paper IV). On the signs, passers-by were asked to write the observed WL-class onto a form and to leave the form in a letterbox. These stations are hereafter referred to as *pen-and-paper stations* and had between 23 and 202 contributions. An overview of the different locations and the type of water level measurements is given in Table 1 of Paper IV. We assessed the agreement between the observed WL-class data and measured water levels using the Kendall rank correlation coefficient (Kendall, 1990). Even though the water level measurement stations can be considered well maintained, errors in the stage measurements cannot be fully excluded, e.g. Horner et al. (2018) found errors in water level measurements in the order of 4 to 12% at six gauging stations in France. However, such inaccuracies are beyond the scope of this study and therefore the water level measurements are considered to be error-free. We furthermore analysed the contribution times to see whether more observations were submitted during summer vs. winter or during weekend vs. weekdays. We also checked in which percentiles of the measured water levels the crowd-based WL-class observations were made to determine if measurements are also made during high flow conditions.

### 4.3.2 Results

#### Accuracy

WL-class observations made with the CrowdWater app by citizen scientists correspond well with measured water levels (Figure 4.5). Even though the results of such WL-class data are not perfect and class boundaries are often fuzzy, the estimated WL-classes from the app are in good accordance with measured water levels. The observed WL-classes for the pen-and-paper stations did not correspond as well with measured water levels as the contributions for the app spots (Figure 4.6).

#### Which water levels are covered?

Our results furthermore show that our citizen scientists in the app often observed high and low flows. Hence, it can be expected that dedicated citizen scientists make observations even when the weather conditions are rather harsh and that some are ambitious to catch exceptional water levels. The main contributor to the spot at the Alp in Einsiedeln (*stealthreporter*; with 32% of the observations at times when the water level was above the 90th percentile of all water level measurements) stated:

> *"The other day, I left the house again because it rained, to observe some high flows."*

For the pen-and-paper stations there were fewer contributions at high flows but rather more at low flows. This indicates that people contributed more during periods with pleasant weather conditions, probably because they did not deliberately go outdoors to contribute stream observations.

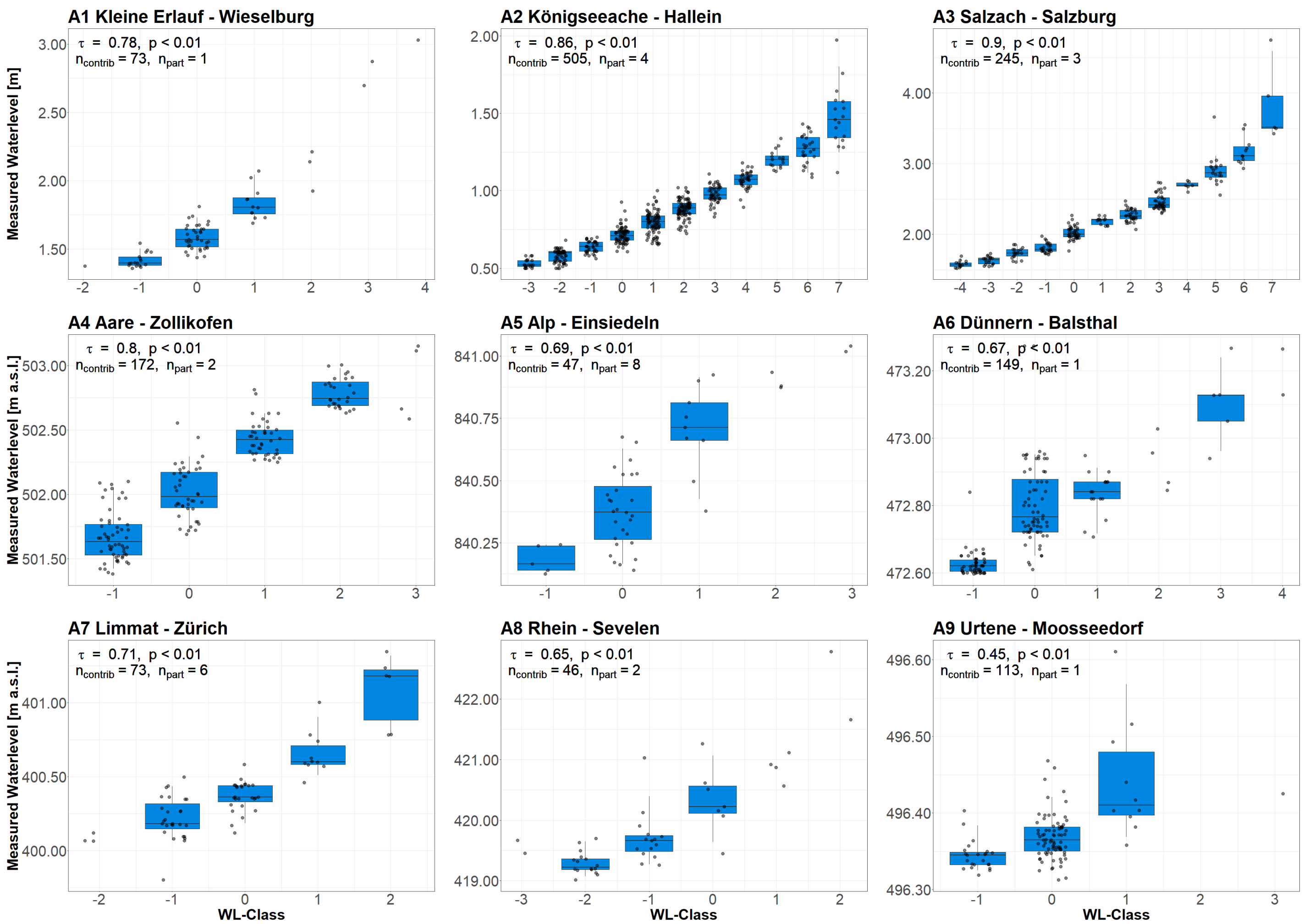



**Figure 4.5:** Correlation between WL-class observations and measured water levels for nine app stations. $\tau$ is the correlation coefficient of the Kendall test, and $p$ the corresponding p-value. $n_{contrib}$ is the number of contributions for the spot and $n_{part}$ the number of individual participants who contributed observations for this spot. The dots are the individual observations and the corresponding measured water level. The boxes show the same data but extend from the $25^{th}$ to the $75^{th}$percentile and the whiskers extend to the $10^{th}$ and $90^{th}$ percentile. The black line inside the box represents the median.

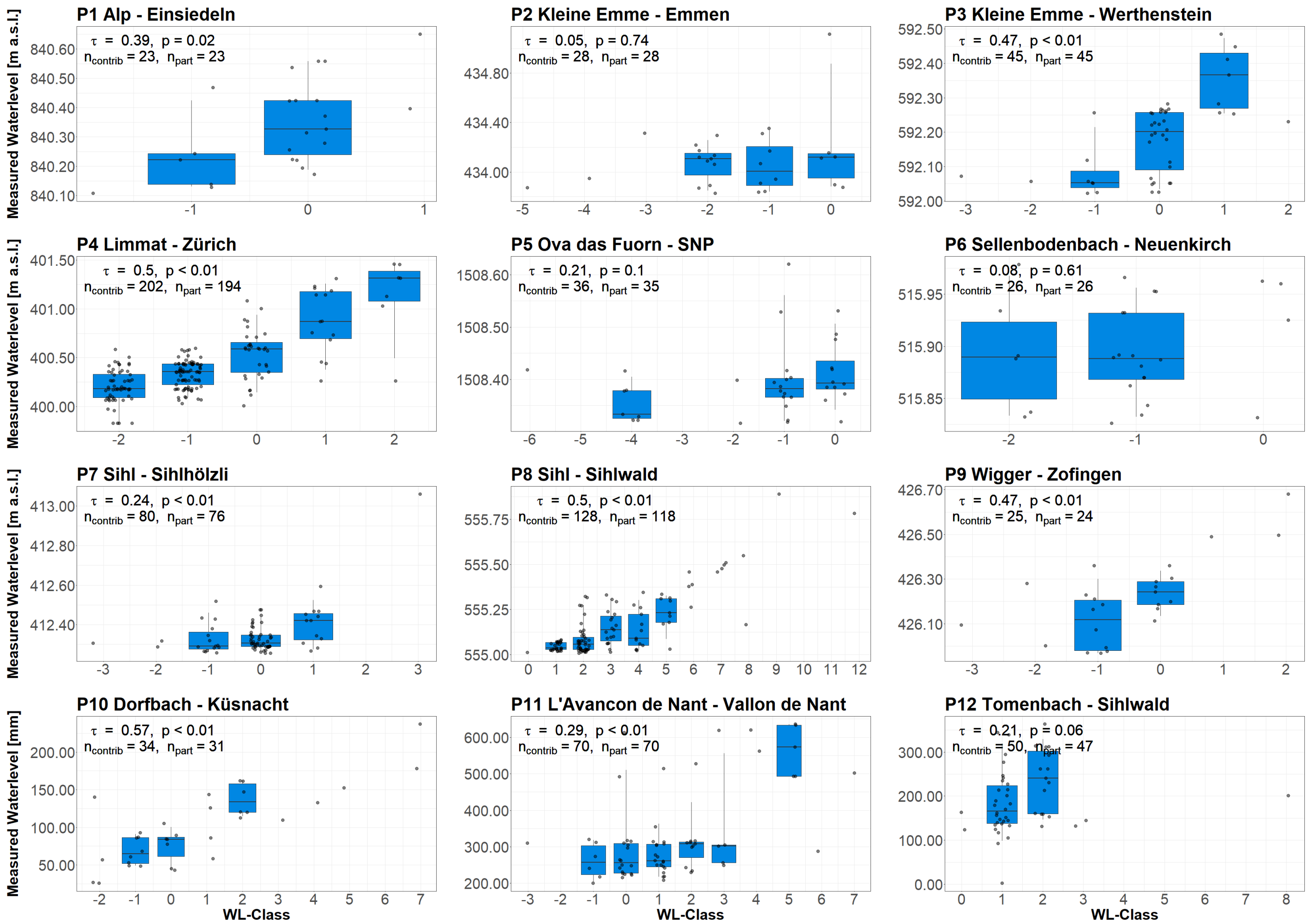



**Figure 4.6:** Correlation between WL-class observations and measured water levels for the pen-and-paper stations. $\tau$ is the correlation coefficient of the Kendall test, and $p$ the corresponding p-value. $n_{contrib}$ is the number of contributions for the station and $n_{part}$ the number of individual participants who contributed to this station. The dots are the individual observations and the corresponding measured water level. The boxes show the same data but extend from the $25^{\text{th}}$ to the $75^{\text{th}}$percentile and the whiskers extend to the $10^{\text{th}}$ and $90^{\text{th}}$ percentile. The black line inside the box represents the median.

**Timing of observations**

Overall, the timing of observations was surprisingly equally distributed throughout the times of the day, days of the week and the months (Figure 4.7). During the days the contributions were rather focused on the early afternoon (pen-and-paper stations) and the late afternoons (app-stations). The pen-and-paper stations had a higher percentage of contributions on weekends, especially on Sundays compared to the app stations where the contributions were more equally distributed throughout the week. We assume that the contributions to the pen-and-paper stations were not part of the daily routines of the citizen scientists but rather occurred when people passed by the stations by chance (during e.g. a walk). Sundays, apparently, are the most likely days for people to be on such walks or hikes. Maybe this is due to the fact, that in Switzerland most shops are closed on Sundays and doing groceries, or other every-day activities is not possible then. Amongst the citizen scientists who used the app, there were probably more committed people that planned their contributions as a part of their daily or weekly routine. However, the contribution patterns vary across the spots (Figures S1 and S2 in Paper IV), therefore it becomes unpredictable when a citizen scientist will contribute without knowing more about their daily routines. Throughout the year there was only a slight tendency for more contributions during the warmer months at both the pen-and-paper and the app stations.

## 4.4 Conclusions and implications

The survey results showed that WL-classes are a suitable quantity to be estimated by citizen scientists. The results also showed that the accuracy of streamflow estimates was lower than the accuracy of WL-class estimates and that variations in the flow conditions were not fully discernible in the streamflow estimates. In addition to being more accurate than streamflow estimates, the WL-class estimation process is also very quick, which is a big advantage for a citizen science project. It is assumed that offering a fast procedure to document stream levels will encourage citizen observers to contribute data to a project regularly (Eveleigh et al., 2014).

The results from Paper IV showed that citizen scientists can collect time series of WL-classes that are in good accordance with measured water levels with the CrowdWater app and the virtual staff gauge approach. Observations with the CrowdWater app lead to better results compared to the pen-and-paper stations. We assume that the lower data quality for the pen-and-paper stations is related to the number of contributors and their familiarity with the method. In the app, the data for each spot was mainly submitted by a single person, whereas for the pen-and-paper station almost every contribution was made by a new participant. Because the virtual staff gauge method largely depends on human perception, different people might come to different conclusions for the same reference image. We assume that our approach with the virtual staff gauge is harder to understand than the approach of CrowdHydrology (Lowry et al., 2019) or of the project in Kenia by Weeser et al. (2018) where water levels in e.g. centimetres are read from phys-

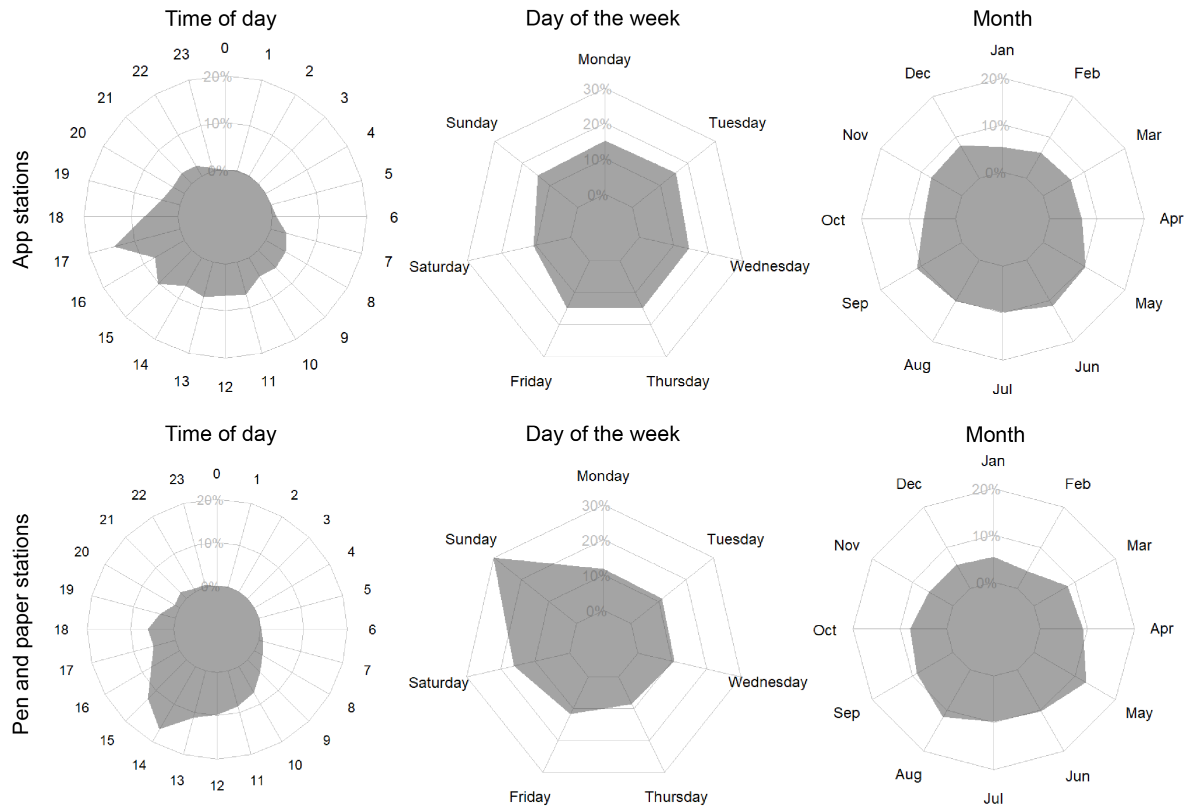



**Figure 4.7:** The grey areas indicate the average percentage of contributions made at all app spots and pen-and-paper stations in this study per time of the day (left), day of the week (middle), and per month (right). The plots for all individual stations can be found in the supplemental material of Paper IV in the Figures S1 and S2.

ical staff gauges. Therefore it might be beneficial for the virtual staff gauge approach, if a few (or a single) dedicated contributors collect all the data. Even if a single contributor had a bias (e.g. always estimating too low) this would result in a more consistent time series than if many people with different biases and perceptions contributed observations to the same station. It is very common for citizen science projects that the majority of the contributions come from a small group of dedicated contributors (Eveleigh et al., 2014; Lowry & Fienen, 2013; Sauermann & Franzoni, 2015). For example, in the Crowd-Hydrology project, one participant walked past a particular station three to four times a week, which led to this station having almost 10 times as many measurements as the station with the next highest number of data submissions (Lowry & Fienen, 2013). This highlights the extreme value of these dedicated contributors.

Another approach would be, that people at the pen-and-paper stations submit only a photo by e-mail using their smartphones and then the WL-class could be estimated by a collective effort, as e.g. in the CrowdWater-Game (Strobl et al., 2019). For the pen-and-paper and for the app approach, increased interaction with the local population might help to improve participation rates of individuals (Lowry et al., 2019). Potentially errors could be reduced through training and information events. Loiselle et al. (2016) found that citizen scientists who got to choose the site at which they contributed data to the FreshWaterWatch project made more repeated measurements compared to participants who were assigned to a station. Furthermore, they also found that if many people contributed to the same stations, then the number of contributions by a single contributor were smaller. This might to some extent be applicable to our setup as well. People who see a sign by chance and decide to contribute feel less committed than those who actively decide to contribute and setup their own observation locations in the app. We assume that creating and maintaining own spots fosters feelings of autonomy and competence, which are in combination with the relatedness of one's own contributions to a broader topic, the basic principles of self-determination theory (Deci & Ryan, 2000). The theory says that the motivation to participate increases, the more the desire for autonomy, competence and relatedness are fulfilled (Frensley et al., 2017). This might explain the favourable behaviour of the citizen scientist who was so motivated to observe high flows that he went out deliberately to do so when it rained.

The errors in the WL-class estimates could be smaller if the participants of the pen-and-paper stations would have undergone some form of training e.g. with the CrowdWater-Game (Strobl et al., 2019) or would have contributed multiple times to gather more experience.

The results are encouraging for using citizen science in hydrology and demonstrate that using a smartphone application for crowd-based WL-class observations is a promising approach. These findings provide an empirical basis to quantify the accuracy of CrowdWater estimates and formed the basis for evaluating the potential value streamflow and WL-class observations for hydrological modelling (Paper V and Paper VI). It remains to be investigated in what ways these CrowdWater timeseries of WL-classes have the potential to complement traditionally measured streamflow time series besides their use in hydrological models to obtain simulated streamflow (see Paper VI).

**Conclusions and implications for CrowdWater**

The findings of Paper I, Paper III, Paper IV and the literature provide insights on beneficial practices that increase the quality of CrowdWater spots and the value of the resulting WL-class time series:

- The reference image with the virtual staff gauge should be taken preferably at low flow, as more features remain visible that facilitate a comparison with reality for the subsequent observations (Paper I).
- Distinct features in the reference image are necessary to identify changes. Vegetation can hinder a clear identification of these features (Paper I).
- The picture needs to be taken as level with the water surface as possible to avoid distortion of the view (Paper III).
- Shorter distances between the observer and the location of the virtual staff gauge and the reference structures have a positive impact on the quality of WL-class estimates. On wider rivers it is therefore beneficial to use features in the stream as e.g. bridge pillars (Paper III).
- Staff gauge size needs to be appropriately sized for water level fluctuations to catch most of the variability as can be seen in e.g. the station Salzach Salzburg compared to Urtene, Moosseedorf (Figures 2 and 7 in Paper IV).
- Dedicated citizen scientists are needed to obtain data for a range of flow conditions (Paper IV).
- Feedback and visibility of participants' contributions might help sustained participation (Lowry et al., 2019). We assume that the app to some extent fulfils these criteria by displaying all the contributions publicly compared to the simple pen-and-paper stations. However, feedback on how the data are used and what individual contributions add to scientific research needs to be communicated as well (Eveleigh et al., 2014).

# 5

# What is the value of crowd-based streamflow, water level and WL-class data for hydrological model calibration?

## 5.1 Introduction

Hydrological models are important tools to study the impacts of natural and anthropogenic changes in a catchment. They can, furthermore, be used in water management and for flood or drought forecasting. The application of such models usually requires several years of precipitation, temperature and streamflow data for calibration, but these data are only available for a limited number of catchments. Therefore, several studies have addressed the question: how much data are needed to calibrate a model for a catchment? Many of them concluded that a limited number of streamflow measurements can be informative to sufficiently calibrate a hydrological model (Brath et al., 2004; Juston et al., 2009; Perrin et al., 2007; Pool et al., 2017; Seibert & Beven, 2009; Seibert & McDonnell, 2015). Seibert & Vis (2016) and van Meerveld et al. (2017) investigated the potential of water level and WL-class data respectively for hydrological model calibration. They found that water level data was informative for model calibration, especially in humid catchments (Seibert & Vis, 2016) and that WL-class data also led to a better model performance than model runs using random parameter sets (i.e., lower benchmark, representing a situation without any data). Although the above studies had different foci and used different model performance metrics their results are nevertheless encouraging for the calibration of hydrological models for ungauged basins based on a limited number of crowd-based streamflow, water level or WL-class observations. One aim of the CrowdWater project is to continue this line of research and to develop a methodology that allows citizen scientists to collect data that is informative for hydrological model calibration.

This chapter therefore summarises Paper V in which we investigated the potential value of crowd-based streamflow estimates and Paper VI where we tested the potential value of water level measurements and WL-class estimates. A simplified graphical summary of the two papers is given in Figures 5.1 and 5.2

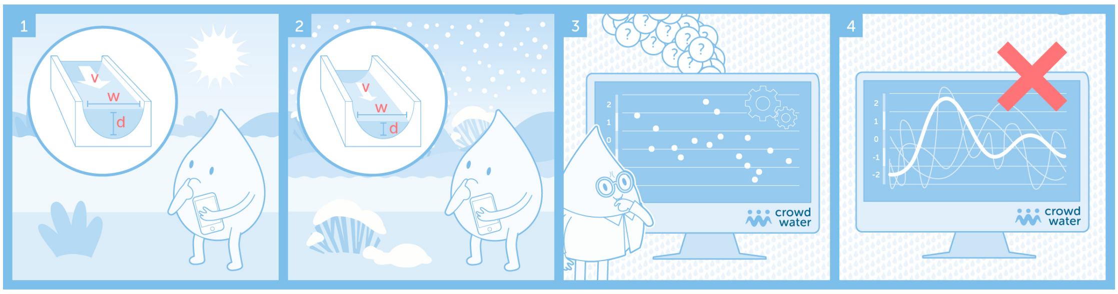



**Figure 5.1:** A simplified illustration of the findings in Paper V that depicts the streamflow estimation via stream width, average stream depth and the flow velocity. In the last image shows that the uncertainty in these estimates is too high to be directly informative for hydrological model calibration. Design by: University of Zurich, Information Technology, MELS/SIVIC, Tara von Grebel


**Figure 5.2:** A simplified illustration of the findings in Paper VI that depicts the water level class estimation with the virtual staff gauge. The last image shows that such water level class estimates are informative for hydrological model calibration. Design by: University of Zurich, Information Technology, MELS/SIVIC, Tara von Grebel

## 5.2 Methods

### 5.2.1 HBV-light model

The bucket-type, semi-distributed hydrological model HBV (Hydrologiska Byråns Vattenavdelning; Lindström et al. 1997) was originally developed at the Swedish Meteorological and Hydrological Institute (SMHI) by Bergström (1976). We used the version HBV-light (Seibert & Vis, 2012). In this section the model variant, routines and parameters (denoted by a leading $P$) that were used for Paper V and Paper VI are explained (Table 5.1 and Figure 5.3). We used time series of measured precipitation, temperature and potential evaporation (PE) with hourly resolution as input data. Elevation zones (each 200 $m$) allowed representation of the increase in precipitation (via the gradient-parameter $P_{PCALT}$ [$\%100\,m^{-1}$]), and the decrease in temperature (via the gradient-parameter $P_{TCALT}$ [$°C100\,m^{-1}$]) with increasing elevation. We did not use any vegetation zones or different aspects of the elevation zones. This separation allows to treat precipitation as either rain or snow based on the adapted temperature in each elevation zone. If the temperature was below the temperature threshold $P_{TT}$ [$°C$], the precipitation was considered to be snow and was corrected by the snowfall correction factor $P_{SFCF}$ [$-$] to account for systematic errors in snow measurements and the evaporation losses from the snow pack, which are not explicitly modelled. Snowmelt in each elevation zone was calculated using a degree-day-factor $P_{CFMAX}$ [$mm°C^{-1}h^{-1}$] (equation 5.1):

$$snowmelt = P_{CFMAX}(T(t) - P_{TT}) \tag{5.1}$$

where $T(t)$ was the temperature at each time step and $P_{TT}$ the threshold for melt to occur. Meltwater and rainfall are stored within the snowpack up to the exceedance of the fraction $P_{CWH}$ [$-$], which is the maximum water equivalent of the snowpack. If the temperature of the timestep $\Delta t$ was below $P_{TT}$, the refreezing in the snowpack was calculated using equation 5.2:

$$refreezing = P_{CFR} * P_{CFMAX}(P_{TT} - T(t)) \tag{5.2}$$

where $P_{CFR}$ [$-$] is the coefficient of refreezing.

The sum of the liquid precipitation and melt water are either input $I(t)$ to the soil box of the corresponding elevation zone or are directly recharging $R(t)$ the groundwater in the upper groundwater box $S_{UZ}$ (Figure 5.3). This fraction depends on the previous water content of the soil box $S_{SOIL}(t)$ and its largest possible value $P_{FC}$ [$mm$] (equation 5.3):

$$\frac{R(t)}{I(t)} = \left(\frac{S_{SOIL}(t)}{P_{FC}}\right)^{P_{BETA}} \tag{5.3}$$

where $P_{BETA}$ determines the relative contribution to runoff from rain and snowmelt. The groundwater boxes $S_{UZ}$ and $S_{LZ}$ are lumped, i.e. there is only one box for the entire catchment (Figure 5.3). The actual evaporation $AE$ equalled $PE$ if the water content of the soil box divided by $P_{FC}$ is above $P_{FP} * P_{LP}$ [$-$], else a linear reduction is used (equation 5.4):

$$AE = PE(t) * min\left(\frac{S_{SOIL}(t)}{P_{FC} * P_{LP}}, 1\right) \tag{5.4}$$

where $P_{LP}$ [$-$] is a threshold for the reduction of evaporation. $P_{PERC}$ [$mmh^{-1}$] defines the percolation of the upper groundwater box to the lower groundwater box (Figure 5.3).

Runoff $Q$ from the groundwater boxes is computed as the sum of the three linear outflow equations depending on whether $S_{UZ}$ is above a threshold value, $P_{UZL}$ [$mm$], or not (equation 5.5).

$$Q_{1+2+3} = P_{K2} * S_{LZ} + P_{K1} * S_{UZ} + P_{K0} * max(S_{UZ} - P_{UZL}, 0) \tag{5.5}$$

**Table 5.1:** Description of the HBV-light parameters. Seibert & Vis (2012) and their ranges used for calibration of the model.

| Parameter | Description | Unit | Min | Max |
|---|---|---|---|---|
| *Rescaling Parameters of Input Data* | | | | |
| PCALT | change in precipitation with elevation | $\% \ 100m^{-1}$ | 5 | 15 |
| TCALT | change in temperature with elevation | $°C \ 100m^{-1}$ | 0.5 | 1.5 |
| *Snow and ice melt parameters* | | | | |
| TT | threshold temperature for liquid and solid precipitation | $°C$ | -3 | 1 |
| CFMAX | degree-day factor | $mm \ °C^{-1} \ h^{-1}$ | 0.06 | 10 |
| SFCF | snowfall correction factor | $-$ | 0.4 | 1.6 |
| CFR | refreezing coefficient | $-$ | 0.001 | 0.9 |
| CWH | water holding capacity of the snow storage | $-$ | 0.001 | 0.9 |
| *Soil Parameters* | | | | |
| PERC | maximum percolation from upper to lower groundwater storage | $mm \ h^{-1}$ | 0 | 3 |
| UZL | threshold parameter | $mm$ | 0 | 100 |
| K0 | storage (or recession) coefficient 0 | $h^{-1}$ | 0.001 | 0.5 |
| K1 | storage (or recession) coefficient 1 | $h^{-1}$ | 0.0001 | 0.2 |
| K2 | storage (or recession) coefficient 2 | $h^{-1}$ | 2.00E-06 | 0.005 |
| MAXBAS | length of triangular weighting function | $h$ | 1 | 7 |
| FC | maximum soil moisture storage | $mm$ | 50 | 550 |
| LP | soil moisture value above which actual evapotranspiration $AE$ reaches potential evapotranspiration $PE$ | - | 0.3 | 1 |
| BETA | shape factor for the function used to calculate the distribution of rain and snow melt being routed to the soil box ($S_{SOIL}$) or the groundwater ($S_{UZ}$), *respectively* | - | 1 | 5 |

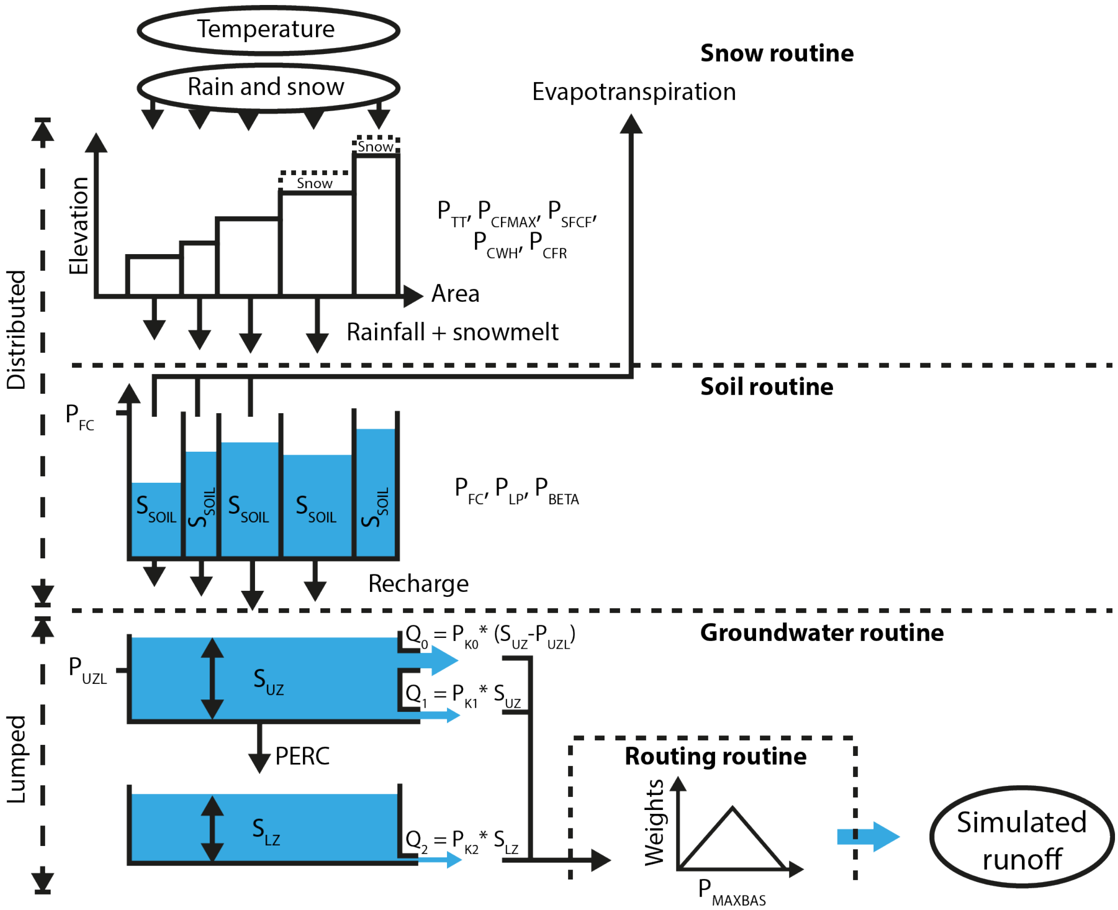



**Figure 5.3:** The structure of the HBV-light model (adapted from Uhlenbrook et al., 1998).

This outflow is then transformed by a triangular weighting function that is governed by $P_{MAXBAS}$ $[-]$ (equation 5.6) and results in the simulated streamflow $Q_{sim}$ for each time step in $[mm\,h^{-1}]$.

$$Q_{sim}(t) = \sum_{i=1}^{P_{MAXBAS}} c(i) * Q_{1+2+3}(t-i+1),$$
$$\text{where } c(i) = \int_{i=1}^{i} \frac{2}{P_{MAXBAS}} - \left| u - \frac{P_{MAXBAS}}{2} \right| * \frac{4}{P_{MAXBAS}^2} du \tag{5.6}$$

### 5.2.2 Data

In Paper V, we calibrated the HBV-model for six catchments in Switzerland. For Paper VI, we used only four of these catchments because the performance of Allenbach and Riale di Calneggia was bad due to issues with the rainfall and/or streamflow data, which led to rainfall-runoff ratios >1 (the detailed rainfall-runoff ratios and catchment characteristics can be found in Table 2 of Paper V, Table 1 of Paper VI and in Figure 5.4). All streamflow and water level data were obtained from the FOEN. All rainfall and temperature data were obtained from MeteoSwiss. For each of the catchments we selected a dry, a wet and an average year within the period 2006-2014 based on the total summer streamflow for model calibration and validation. The years were the same in both studies.

### 5.2.3 Creation of synthetic datasets

We fitted a continuous normal distribution to the logarithms of the streamflow estimates relative to the measured streamflow (i.e., error distribution) for the medium sized streams (Töss, Sihl and Schanzengraben in the Canton of Zurich and the Magliasina in Ticino; n=136) from Paper III (Figure 4.4(a)). These medium sized streams had a similar streamflow range at the time of the estimations of 2.6 – 28 $m^3/s$ as the mean annual streamflow of 1.2 – 10.8 $m^3/s$ of the streams in the six catchments used for model calibration in Paper V. We used this error distribution for streamflow together with the observed streamflow time series to generate synthetic streamflow series that represent the uncertainties of real crowd-based estimates for the six catchments of Paper V. This representation of streamflow estimation accuracy allowed us to generate an uncertain streamflow or water level class observation for every time step of the measured streamflow data based on an empirically based probability.

Similarly, for the creation of synthetic WL-class time series we used a discontinuous normal distribution that we fitted to the class errors for the WL-class estimates determined by us in Paper III (Figure 4.4(b)). For the creation of synthetic WL-class time series we used time series of the measured water levels that we binned into 2-10, 15, and 20 classes. We then generated random noise with the magnitudes and associated likelihoods from the normal distribution for every time step on the WL-class time series with 10 classes. As a result, 48% of all WL-class observation points were correct, roughly

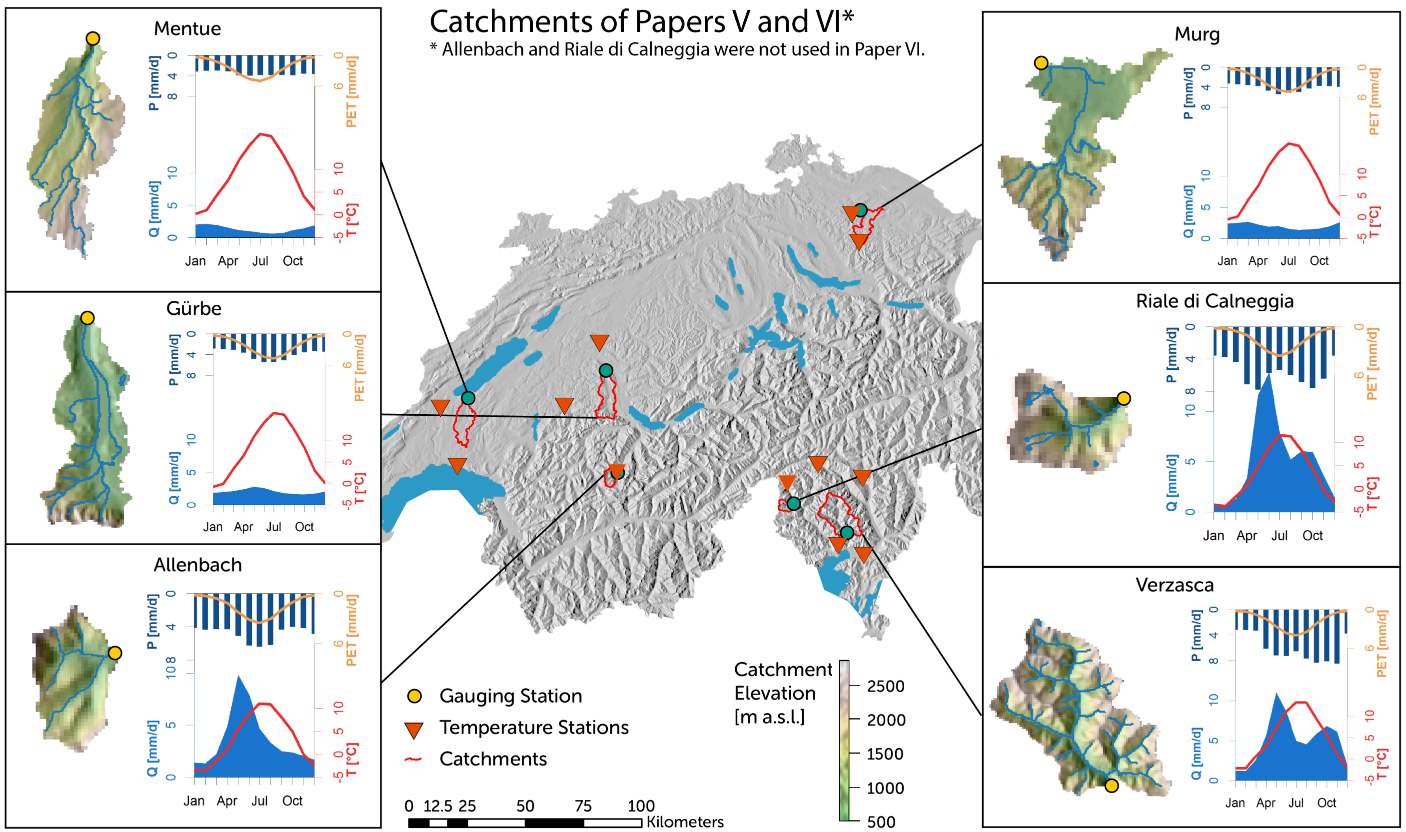



**Figure 5.4:** The catchments that were used in Paper V and Paper VI and their locations within Switzerland. The inset graphs show the mean monthly precipitation (P), streamflow (Q), potential evaporation (PE) and temperature (T). The catchments Allenbach and Riale di Calneggia were not used in Paper VI.

40% of all classes were one class higher or lower than the correct class and roughly 13% of the data points were more than one class off.

For both the streamflow and WL-class time series, we also generated time series with smaller errors based on the same error distribution but with the standard deviation divided by two and four. For every class and error magnitude, we then created time series with fewer data points reflecting different scenarios of likely contribution times. This resulted in uncertain streamflow and WL-class time series with *Hourly*, *Weekly*, *Daily*, *Monthly* observations, two time series with measurements during weekends in the period from March to August *WeekendSpring* or from May to October *WeekendSummer* and every other day during the months of July, August, and September *IntenseSummer*. Furthermore, we generated a scenario with 52 (*Crowd52*) and 12 (*Crowd12*) data points per year, with a higher probability for contributions at times when we assumed that people were more likely to be outdoors (i.e. most contributions in summer, only during daylight, and outside working hours).

### 5.2.4 Model calibration and validation

#### Calibration procedure

For all the model calibrations with measured and synthetic streamflow, we used the overall performance index ($P_{OA}$; Finger et al., 2011). The $P_{OA}$ is the mean of the Nash-Sutcliffe efficiency for the streamflow(Nash & Sutcliffe, 1970), the Nash-Sutcliffe efficiency for the log-transformed streamflow, the mean absolute relative error, and the volume error. For each calibration with water levels or WL-classes, we optimized the Spearman rank correlation coefficient (Spearman, 1904) between the synthetic WL-class data set and the simulated streamflow using a genetic optimization algorithm (Seibert, 2000). The advantage of using the Spearman rank correlation is that it does not require any information on the rating curve for calibration based on water levels or WL-classes. A good fit is obtained as along as simulated streamflow and observed water levels or WL-classes go up and down simultaneously and therefore the dynamics are the same. The assumption is, that the water balance is largely constrained by the precipitation inputs (Seibert & Vis, 2016). To consider parameter uncertainty, the calibration was performed 100 times, which resulted in 100 parameter sets for each case. The parameter sets and their ranges used for calibration can be found in Table 5.1. For each case, the preceding year was used for the warm-up period. For the *Crowd52* and *Crowd12* time series, we used 100 different random selections of times, whereas for the regularly spaced time series the same times were used for each of the 100 calibrations. For the synthetic streamflow data this resulted in a total number of 576 calibrations (6 catchments, 3 calibration years, 4 error groups, 8 temporal resolutions) and for the synthetic WL-classes the total number of calibrations was 3'564 (4 catchments, 3 calibration years, 9 different temporal resolutions, 3 error magnitudes with 10 classes, and 11 class sizes without errors.

**Validation procedure**

The model validation for all cases was performed using the $P_{OA}$ based on the obtained parameter sets from the calibration. The obtained parameter sets from the different year characters were cross-validated with all three year characters each. The validation performance of the model calibrated with one year of measured streamflow data served as the upper benchmark (Seibert et al., 2018) and represented the best possible situation with high resolution and high quality streamflow data. For the lower benchmark we used 1000 randomly generated parameter sets, which represented a situation were no streamflow data was available for model calibration.

## 5.3 Results

The results for the streamflow estimates and WL-class estimates differed: the effect of errors was greater in the streamflow scenarios than for the WL-class scenarios (Figure 5.5). The impact of typical errors for citizen-science-based estimates of WL-classes on the model performance was small. This is perhaps not surprising, as about half of the WL-class estimates were still correct in the scenario with the largest errors. The errors in streamflow data had a much larger impact because they were often larger than the natural fluctuations in streamflow. These results indicate that streamflow estimates from untrained citizens are not directly informative for model calibration. However, if the errors are reduced, the estimates are informative and useful for model calibration (see difference between large and medium errors in Figure 5.5). As expected, the model performance increased when the number of streamflow estimates used for calibration increased. The model performance was also better when the streamflow estimates were more evenly distributed throughout the year (Figure 5.5).

The results of the calibrations using WL-class data in Paper VI indicate that on average one WL-class observation per week for a one-year period (see Crowd52 scenario) can significantly improve model performance compared to the situation without any streamflow data. In fact, the validation performance for model parameters calibrated with 52 WL-class observations was similar to the performance of the calibration with precise water level measurements (as can be obtained from a water level logger; see comparison water levels and WL-classes in Figure 5.5). Errors in the estimates (Figure 5.5) and the number of WL-classes (when at least four to five WL-classes were used) did not influence the validation performance noticeably (Figure 5.6).

Although there was a general trend of increasing model performance with an increasing number of observations, the timing of the observations within the year also had a substantial effect on model performance. The validation performance for the model calibrated with *Crowd52* data (i.e., with more observations in summer) was comparable to the performance of the model calibrated with *Hourly* water level data, regardless of the number of classes. On the other hand, the model validation performance of the model calibrated with *Weekly* data was significantly worse than the performance of the model calibrated with *Hourly* water level data, even when using 20 WL-classes. This is contrary to the results for uncertain streamflow observations of Paper V, where *Weekly* data resulted in a better model validation performance than *Crowd52* data. For WL-class estimates, it is probably beneficial to obtain observations that cover a larger variation in streamflow magnitudes than for streamflow directly because it takes a relatively large change in the actual water level (and thus also streamflow) to change one WL-class.

The results of Paper V indicate that streamflow estimates from untrained citizens are not directly informative for model calibration. However, if the errors could be reduced, the estimates are informative and useful for model calibration. As expected, the model performance increased when the number of observations used for calibration increased. The model performance was also better when the observations were more evenly distributed throughout the year. However, the results of Paper VI indicate that on average one WL-class observation per week for a one-year period (see *Crowd52* scenario) can significantly improve model performance compared to the situation without any streamflow data. Furthermore, the validation performance for model parameters calibrated with WL-class observations was similar to the performance of the calibration with precise water level measurements. The number of WL-classes did not influence the validation performance noticeably when at least four WL-classes were used. The impact of typical errors for citizen-science-based estimates of WL-classes on the model performance was small.

## 5.4 Conclusions and implications

The results of WL-class simulations from Paper VI are encouraging for citizen science projects because they suggest that the observations of water levels by citizens using virtual or physical staff gauges for otherwise ungauged streams provide useful information for model calibration. Although the validation performance of the model calibrated with synthetic WL-class data with realistic frequencies for citizen science projects was not as good as when streamflow data were used for calibration, the performance was comparable to a calibration with data collected with water level loggers or physical staff gauges with precise markings. Because the results of Paper VI showed, that collecting WL-class data at different magnitudes of streamflow is beneficial for model calibration, there might be a concern, that citizen scientists would only contribute data during favourable weather conditions, and thus not make high flow observations. However, based on the results of Paper IV it is realistic to expect that that citizen scientists also contribute during high flows, even when the weather conditions are harsh. The WL-class observation approach has the advantage of being easier to implement and more scalable because it does not require any physical installations (and, thus, no special equipment, permits or maintenance). We can therefore conclude that crowd-based WL-class time series and also more precise water level time series (Weeser et al., 2019) can be useful to inform hydrological models in regions where otherwise no data would be available, if on average at least one observation is made per week for one year. Contrary, the results from the calibration with synthetic streamflow estimates (Paper V) suggest that it is not useful to have citizens estimate streamflows, unless their errors can be reduced by training. This

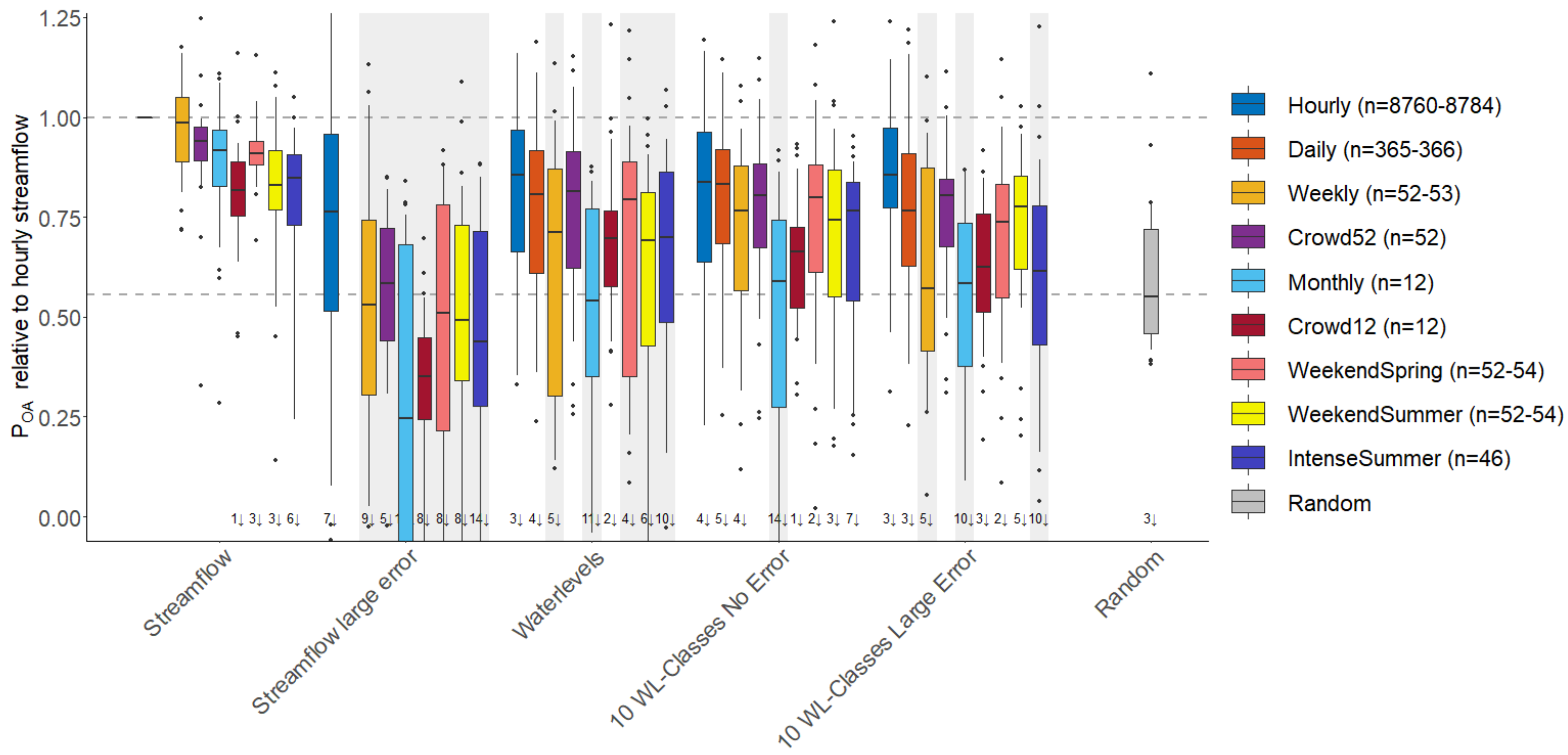



**Figure 5.5:** Box plots of the model validation performance of the HBV-model calibrated with the data of Paper V and Paper VI water level data with different temporal resolutions and the synthetic WL-class data (ten classes) with different temporal resolutions and different errors, relative to the validation performance of the model calibrated with hourly streamflow data (upper benchmark). The lower benchmark shown (in grey) is the median validation performance of the model run with 1000 random parameters. Note that there are no Daily scenarios for the streamflow simulations from Paper V. The grey shading indicates a median model performance that is not significantly better than the lower benchmark ($p>0.05$). The box extends from the $25^{th}$ to the $75^{th}$ percentile and the whiskers extend to the $10^{th}$ and $90^{th}$ percentile. The black line inside the box represents the median. Numbers at the bottom indicate outliers with a relative $P_{OA}<0.00$. Note that the boxes for the calibrations with streamflow are not the same as in Paper V because the data of the catchments Allenbach and Riale di Calnegia were removed as they were not used for the calibrations with water levels and WL-classes.

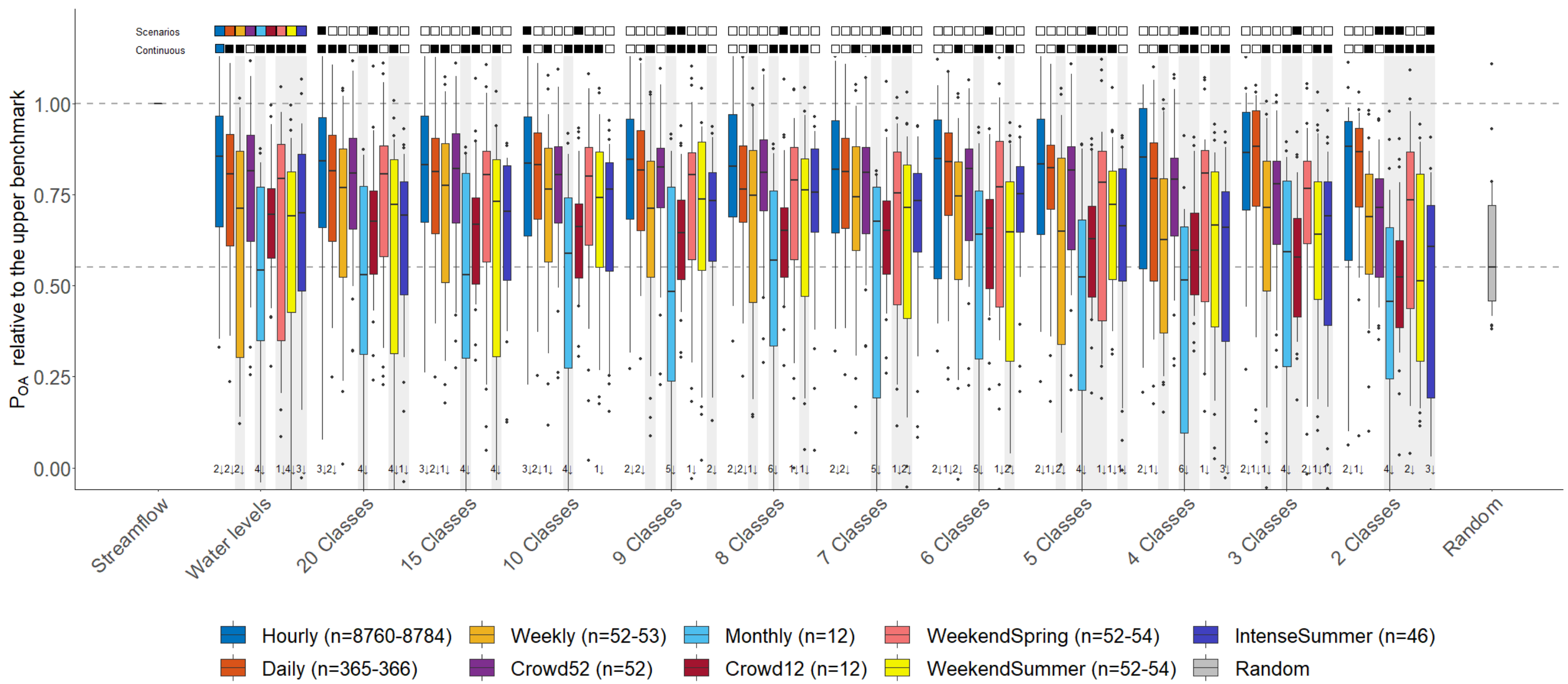



**Figure 5.6:** Box plots of the validation performance of the HBV-model calibrated with synthetic WL-class data (different temporal resolutions and different numbers of WL-classes) relative to the performance of the model calibrated with hourly streamflow data. The lower benchmark (in grey) represents the median performance of the model run with 1000 randomly selected parameter sets. The grey background shading highlights the scenarios for which the median model performance was not significantly better than for the lower benchmark. The filled squares at the top of the graph indicate cases where the median validation performance for the model calibrated with WL-class data was significantly worse compared to the calibration with water level data with the same temporal resolution (top row) and compared to the calibration with continuous (hourly) water level data (second row); empty squares indicate no statistically significant difference based on the one-sided paired Wilcoxon test. All scenarios led to a significantly worse model validation performance than calibration with continuous streamflow data. The WL-classes were equally sized and assumed to be error free. The box extends from the 25$^{\text{th}}$ to the 75$^{\text{th}}$ percentile and the whiskers extend to the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentile. The black line inside the box represents the median. Numbers at the bottom indicate outliers with a relative $P_{OA}$<0.00. Figure obtained from Paper VI.

suggests that it is more useful to focus the efforts of citizens on observations of WL-class data, and when needed, to use models to convert these estimates into streamflow than to ask them to estimate the streamflow directly.

# 6

# Summary, discussion and suggestions for future research

## 6.1 Motivation of citizen scientists

In Paper II, we showed the CrowdWater and Naturkalender participants mainly joined the projects to contribute to science, to satisfy their interest in science and technology, to protect nature, contribute to the well-being of society, learn something new, and to be physically active. Fun and enjoyment were not the primary motivations to become involved in the projects but were essential motivators for continued participation.

At the time of the survey, about half of the CrowdWater users agreed that social pressure had led to their involvement in the project. This may have changed by now, as many of the people who were active back then probably stopped participating. On the other hand, many new participants joined and the number of participants with at least one contribution has more than doubled since then (265 in October 2018 vs. 585 in January 2020). We assume that the motivations of CrowdWater participants are now more similar the motivations of Naturkalender participants because the participant basis is now dominated by people that we did not know before and that might therefore be more self-motivated. Some CrowdWater participants contribute as part of their job or their research and might, therefore, be motivated by more extrinsic motivations or by pushing their career. The learning aspect, however, did not change in the two projects and therefore the motivations related to learning might not change much. On the other hand, with the new feature to document plastic pollution, more participants with the desire to protect the environment and to help society might join.

## 6.2 Hydrological research

In chapters 4 and 5 of this thesis, I showed that crowd-based WL-class time series obtained with the virtual staff gauge approach can provide useful data in regions where otherwise no data would be available. In Paper I, we showed that the majority of participants understood the concept of the virtual staff gauge. Paper III demonstrated the higher accuracy of WL-class estimates, especially also when compared to streamflow estimates. The results of Paper III allowed us to generate synthetic streamflow time and WL-class time series with the uncertainties that can be expected if citizen scientists make the observations. With these time series, we then investigated the potential of such data for the calibration of hydrological models. The results of Paper V showed that streamflow estimates with uncertainties that are realistic for untrained citizen scientists, do not provide any value compared to a situation without any data. However, if the errors could be reduced, they might be informative for model calibration. The results of Paper VI show that water level- or WL-class observations are useful for model calibration, compared to a situation without any data, if at least one value per week over one year was used for calibration. The models calibrated with water levels or WL-class estimates (even with the largest errors) had a significantly higher validation performance compared to a situation without any data. However, models calibrated with water level or WL-class observations performed significantly worse than models that were calibrated with hourly streamflow data. The results of Paper VI also showed that the benefits of having more than four to five classes were negligible. The virtual staff gauge in the CrowdWater app has ten classes. This allows WL-class estimates to be useful, even if citizen scientists make the staff gauge too big to perfectly cover all water level fluctuations. Too small virtual staff gauges, would not only make it harder to distinguish classes but would also lead to missed information at flows higher or lower than the virtual staff gauge. However, in the past years and to my knowledge, such a case has never occurred in the CrowdWater app.

In Paper VI, we found that WL-class time series that were obtained via the app and were submitted by mainly one person, were more consistent and in better agreement with the measured water level data than the data that were obtained from the letterboxes and were contributed by many different people. This suggests that the errors used in Paper VI are perhaps too large and the median sized errors may be more representative. However, the performance of the model calibrated with the WL-class data was insensitive to the errors in the WL-class data and this is thus not likely to affect the results. However, if the better data quality for estimates from a single person compared to a group of people would also hold for the streamflow estimates, this finding could mean that the quality of the streamflow time series may be better than estimated in Paper III and are perhaps best represented by the results for the medium errors, if streamflow would always be estimated by the same person. The personal estimation accuracy of individuals could maybe even be improved by providing feedback on the accuracy of their estimates, if estimates are made in the vicinity of gauging stations. In the app, the option to estimate streamflow exists, but it is not promoted and hardly ever used. Therefore, we have no such time series and cannot confirm this assumption. If there is a constant bias for an individual observer, one could also use Spearman ranks to calibrate the model, instead of the common objective functions for calibration with streamflow like, e.g. the Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970).

## 6.3 Recommendations

### 6.3.1 Future research directions

The potential of crowd-based data for model calibration could be studied for different characteristics of the hydrograph, such as the timing of peaks, the representation of the overall water balance or the simulation of high and low flows. This could be done by comparing validation objective functions that describe the model performance for one of these characteristics (e.g. the logarithmic Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) for low flows).

The logical next step is to calibrate and validate hydrological models with real crowd-based WL-class time series. Thereby further potential uncertainties could come into play, which were not considered in Paper V and Paper VI like the placement of the virtual staff gauge, the suitability of a location, the dynamics of the stream, timing of observations etc. Then also the potential of real crowd-based WL-class data for streamflow forecasting should be examined. This would answer the question if real crowd-based WL-class time series from the CrowdWater project are useful for water management and natural hazard applications, such as flood or drought forecasting or hydropower production. If such a study would show that streamflow predictions are possible based on this data, the same approach could be applied in regions where otherwise no data would be available. In that case, from a scientific point of view, there should be studies that do a cost-benefit and effort-benefit analysis that compares WL-class observations and streamflow measurements. The CrowdWater approach in combination with the app can then promoted as a tool to facilitate data collection for interested stakeholders. If the potential value of the CrowdWater-data for streamflow forecasting in regions without or very little streamflow data is high, then it would become a valuable tool for agencies that are operating with a low budget as well. Furthermore it could also be a valuable tool for many grass-root movements (Seyfang & Smith, 2007) that wish to document, for instance, unauthorized water withdrawals and or plastic pollution. This could either be orchestrated by non-governmental or local organisations, either without the help of scientists or in close collaboration with them. An example is the "Extreme Citizen Science Group" at the University College in London that collaborates with marginalised groups to identify local problems and helps to solve them by combining local and scientific knowledge (Matthias et al., 2014). The app, therefore, has a the potential to become a valuable tool for different projects that cover many of the different models of citizen engagement of Serrano Sanz et al. (2014).

### 6.3.2 CrowdWater app and management

The CrowdWater project had a successful start during the duration of my and Barbara Strobl's PhD. The technical backbone of the project is clearly the CrowdWater app. Although the app offers already many helpful features, such as offline maps, liking, sharing, flagging, following, push messages as well as checking and locking approved contributions, the app and the admin-interface could still be improved. In particular, I suggest that:

- The app incorporates the already planned extensions to enable the entry of actual water levels from physical staff gauges, similar to the CrowdHydrology project (Lowry et al., 2019) and the planned water quality feature.

- It would be beneficial if the push-messages could be sent to specific user groups. These user groups could be defined by geographic region or alternatively, users could subscribe to updates of a specific region or topic to get only the information that is relevant and interesting to them. Then people could be informed about local events, or interesting conditions that would be useful to observe.

- It is possible to define point locations or regions with a circle for certain events. From a hydrology perspective, it might be beneficial to have the option to use more complex geographic boundaries to e.g. map catchments where data is required by using shapefiles.

- It would be helpful if it was possible to check & lock entire time series based on the quality of the root spot (i.e. a well-placed virtual staff gauge in the case of water level). The check & lock feature, so far, can only be applied to individual contributions. Therefore, the green tick is no longer visible on the map after a new observation has been uploaded.

- Adding learning opportunities might increase the motivation of citizen scientists, as was shown in Paper II. Potential ideas are informative pop-ups (e.g. after each contribution) with some facts about water or a stronger link to existing online learning opportunities, such as the open online course on Water in Switzerland[1].

- Once there are studies that prove that the virtual staff gauge approach leads to valuable data that can be used for e.g. streamflow forecasts, the collaboration with local community groups should be promoted. Once an organisations accepts to use the CrowdWater app, more rights should be given to one or multiple local admins, to distribute the workload of the quality control and to give the groups more autonomy and thereby also to enhance their motivation according to the self-determination theory (Ryan & Deci, 2000).

[1] `https://edu-exchange.uzh.ch/courses/course-v1:UZH+Wasser_CH+2019_T1/` (accessed: 09.01.202)

- The images of the CrowdWater game provide a valuable resource for machine learning approaches, e.g. to classify the images automatically. A machine learning model would, however, have to be trained for each CrowdWater spot individually and a large number of classified images would be necessary (i.e. >100 per class). An alternative would also be to combine the classes from the app into larger groups and e.g. distinguish only between low, normal and high flow. For such a model, less images would probably be sufficient. Once such a model is trained and validated, citizen scientists would only need to upload images and the classification step could be dropped but it needs to be studied, whether this affects the motivation of citizen scientists to contribute to the project. Alternatively, also automatic cameras could then be used in spots where data is extremely valuable.
- The CrowdWater app can be promoted as a tool for data collection and also verification (in the CrowdWater Game). Therefore it can be used in many other applications such as the Plastic Spotter[2] project in the Netherlands, which already uses the app to collect data on plastic pollution.

The management of the CrowdWater project and its community included many tasks: the quality control of the app contributions, communication with participants, organising outreach activities at science fairs, teaching activities for school classes, collaboration with official agencies to tests potential applications of the app, communication with the winners of the monthly CrowdWater game and sending out prizes to them. The communication of the project was to a large extent done via e-mail and social media. Twitter, Facebook and Instagram proved to be important communication channels for the project. Facebook was very helpful at the start of the project to advertise it within our social networks and to communicate with potential collaborators. Instagram was initially set up for the communication with younger participants but turned out to be rather helpful for the communication with collaborators on plastic pollution, such as Plasticspotter[3]. Twitter was mainly used to communicate results, events or collaborations in our scientific network. E-mail was used to communicate with individual citizen scientists and to send out the newsletter. We, unfortunately, have no record of how efficient the communication channels were for recruiting new participants. I expect that at least a 50% position could be filled with the work that is needed to achieve the full potential of CrowdWater. Tasks could also be expanded to regularly produce more labour intense content, such as videos of project participants but also interviews with citizen scientists. It would, therefore, be useful if funding bodies fund such positions. This would not only increase the visibility of CrowdWater but would also help to promote citizen science and to simplify the access to scientific knowledge via online media, especially for the younger generations.

[2] `https://plasticspotter.nl/en` (accessed: 25.03.2020)
[3] `www.plasticspotter.nl` (accessed: 09.01.202)

# Acknowledgements

Many people contributed to my successful PhD. Some had a substantial part in it, others made it much more enoyable. I would like to thank all of them wholeheartedly:

- Ilja van Meerveld and Jan Seibert initiated the CrowdWater project and hired me and Barbara as their PhDs in the project. I really enjoyed developing this project from the start and it has grown dear to my heart. I enjoyed the diversity of tasks, especially the interactions with citizen scientists and the numerous outreach activities. At times when the number of tasks seemed overwhelming, Jan allowed us to hire other personnel to take some work off our shoulders and encouraged us to focus on the important things. In my experience such a practice cannot be taken for granted. I am really thankful for that! Ilja was the person that invested most hours in reading manuscripts, improving my writing and questioning my methods which was not always easy but often left very little to the reviewers in the journals and led to pretty smooth review processes. The encouraging atmosphere that Jan and Ilja created, allowed me to make many wonderful experiences as they encouraged and enabled me to contribute with my own ideas to shape the research and the CrowdWater project.

- Barbara Strobl was a great colleague and collaborator. The project benefited greatly from her diligent working attitude and her sense of duty. Countless times, she reminded me to do things that I neglected, took tasks off my shoulders and provided valuable feedback to my work. It was furthermore great to have her as a colleague at the same level in the project, for sharing ideas, experiences and sufferings, which made the PhD-journey much more enjoyable.

- I would like to thank all H2K-members for being part of this journey, especially during the countless coffee and lunch breaks as well as the numerous shared ASVZ visits over lunch and the help they provided with the CrowdWater project. I would like to thank: Leonie Kiewiet for the great times and deep conversations we had during our PhD-times, Florian Lustenberger for doing all the filming for CrowdWater, which I enjoyed a lot, Fabian Maier for sharing the passion for mountaineering, which I wish we would have done more often, Ling Wang for keeping the group together with numerous nice birthday cards, Marc Vis for the technical support and the organisation of numerous H2K-events as well as for the fun and informative coffee breaks and after-coffee-e-mails, Milena Perraudin for the creation of the great CrowdWater teaching materials, Nina Brunner for extending that work, Nathalie Stübi and Franziska Schwarzenbach for their enthusiastic support with the CrowdWater project, Sandra Pool for the encouragement to go running and the good conversations, Benjamin Fischer for introducing me to the bike trails on "Züriberg", Rick and Kirsti for all those funny stories, Jana Erdbrügger for the many cookies, Manuela Brunner for being a model of determination and last but not least Anna Sikorska, Daphné Freudiger, Daniel Viviorli, and Maria Staudinger for sharing their data and expertise when needed.

- I would also like to thank Alysha Coppola and Margarita Saft for the great conversations in K76 as well as Ulrich Hanke, Nicolas Ofiti and Tatjana Speckert from the soil science group for being such great office mates.

- I would like to thank Ilona as the most important person in my life, for her love, her time, her patience and her support but also for helping me to progress as a person. Our sometimes crazy adventures during the last four years in mountaineering and other sports, faith and our relationship helped me to focus on the things that matter most to me in life.

- A big thank you also goes to the organisations Cevi Alpin, the climbing club Horn K2H and the Evangelical Church Illnau-Effretikon who strongly influenced my life in the past four years and gave me the chance to become a tour-guide, climbing instructor and a role model for young adults. Special thanks go to Simon Weinreich, Markus Enderli, Evelyne Krauer, Raphael Moser, Remo Bischof, Jakob Zirngast and Ilona, who encouraged me to step out of my comfort zone and to make many wonderful experiences.

- I would like to thank my parents for shaping my personality and providing the financial basis for my education. I would furthermore like to thank my brother and my sister for sharing many wonderful experiences and for supporting me.

- I would like to give thanks to all the citizen scientists for their contributions! Without these volunteered efforts, my time as a PhD would not have been the same, and many of the results presented in this thesis would not have been possible.

- I would also like to thank the University of Zurich for providing the infrastructure and the ASVZ for providing the sport facilities and courses that helped me to stay healthy since the start of my studies in 2009.

- Thanks to the Swiss National Science Foundation and all the Swiss taxpayers for providing the financial basis for my thesis (project 163008, CrowdWater).

# Bibliography

Alender, B. (2016). Understanding volunteer motivations to particpate in citizen science projects: a deeper look at water quality monitoring. *Journal of Science Communication*, *15*(3), 2–19.

Aono, Y., & Omoto, Y. (1993). Variation in the March mean temperature deduced from cherry blossom in Kyoto since the 14th century. *Journal of Agricultural Meteorology*, *48*, 635–638.

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLOS ONE*, *13*(1), e0189327.
URL `http://dx.plos.org/10.1371/journal.pone.0189327`

Batson, C. D., Ahmad, N., & Tsang, J.-A. (2002). Four Motives for Community Involvement. *Journal of Social Issues*, *58*(3), 429–445.
URL `http://doi.wiley.com/10.1111/1540-4560.00269`

Bergeron, T. (1949). The problem of artificial control of rainfall on the globe. Part II: The coastal orographic maxima of precipitation in autumn and winter. *Tellus*, *1*, 15–32.

Bergeron, T. (1960). Operation and results of "Project Pluvius.". In *Physics of Precipitation, Geophys. Monogr., No. 5*, (pp. 152–157). Amer. Geophys. Union.

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments*, vol. 52. Norrköping, Sweden: SMHI Norrköping, Report RH07.

Beza, E., Steinke, J., van Etten, J., Reidsma, P., Fadda, C., Mittra, S., Mathur, P., & Kooistra, L. (2017). What are the prospects for citizen science in agriculture? Evidence from three continents on motivation and mobile telephone use of resource-poor farmers. *PLOS ONE*, *12*(5), e0175700.
URL `http://dx.plos.org/10.1371/journal.pone.0175700`

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. Tech. Rep. July, Center for Advancement of Informal Science Education (CAISE), Washington, D.C.
URL `http://www.birds.cornell.edu/citscitoolkit/publications/CAISE-PPSR-report-2009.pdf`

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. (2013a). Using gamification to inspire new citizen science volunteers. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications - Gamification '13*, (pp. 18–25). Stratford, Ontario, Canada.

Bowser, A., Hansen, D., & Preece, J. (2013b). Gamifying Citizen Science: Lessons and Future Directions. *Workshop on Designing Gamification: Creating Gameful and Playful Experiences*.
URL `http://gamification-research.org/wp-content/uploads/2013/03/Bowser{_}Hansen{_}Preece.pdf`

Brath, A., Montanari, A., & Toth, E. (2004). Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. *Journal of Hydrology*, *291*(3-4), 232–253.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0022169404000344`

Breuer, L., Hiery, N., Kraft, P., Bach, M., Aubert, A. H., & Frede, H.-G. (2015). HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters. *Scientific Reports*, *5*(16503).

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., De Bièvre, B., Bhusal, J., Clark, J., Dewulf, A., Foggin, M., Hannah, D. M., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandeya, B., Paudel, D., Sharma, K., Steenhuis, T., Tilahun, S., Van Hecken, G., & Zhumanova, M. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, *2*(26), 21.
URL `http://journal.frontiersin.org/article/10.3389/feart.2014.00026/abstract`

Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., & Purves, R. (2016). *European Handbook of Crowdsourced Geographic Information*. February. London: Ubiquity Press, 1 ed.
URL `http://www.ubiquitypress.com/site/books/10.5334/bax/`

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, *176*(1-4), 273–91.
URL `http://www.ncbi.nlm.nih.gov/pubmed/20640506`

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147.
URL `http://doi.wiley.com/10.1111/j.1467-9671.2012.01359.x`

Curtis, V. (2015). Motivation to Participate in an Online Citizen Science Game. *Science Communication*, *37*(6), 723–746.
URL `https://doi.org/10.1177/1075547015609322`

Davids, J. C., Rutten, M. M., Shah, R. D. T., Shah, D. N., Devkota, N., Izeboud, P., Pandey, A., & van de Giesen, N. (2018). Quantifying the connections—linkages between land-use and water in the Kathmandu Valley, Nepal. *Environmental Monitoring and Assessment*, *190*(304), 17.
URL `http://link.springer.com/10.1007/s10661-018-6687-2`

Davids, J. C., van de Giesen, N., & Rutten, M. (2017). Continuity vs. the Crowd—Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations. *Environmental Management*, *60*(1), 12–29.
URL `http://link.springer.com/10.1007/s00267-017-0872-x`

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268.
URL `http://www.tandfonline.com/doi/abs/10.1207/S15327965PLI1104{_}01`

Dickerson-Lange, S. E., Eitel, K. B., Dorsey, L., Link, T. E., & Lundquist, J. D. (2016). Challenges and successes in engaging citizen scientists to observe snow cover: From public engagement to an educational collaboration. *Journal of Science Communication*, *15*(1), 1–14.

Domroese, M. C., & Johnson, E. A. (2017). Why watch bees? Motivations of citizen science volunteers in the Great Pollinator Project. *Biological Conservation*, *208*, 40–47.
URL `http://dx.doi.org/10.1016/j.biocon.2016.08.020`

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., & Cox, A. L. (2014). Designing for dabblers and deterring drop-outs in citizen science. In *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 2985–2994). Toronto, ON, Canada: ACM New York, NY, USA.

Field, A., Miles, J., & Field, Z. (2013). *Discovering statistics using R*. Los Angeles: Sage.

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S., & Burlando, P. (2011). The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model. *Water Resources Research*, *47*(7), 14.
URL `http://doi.wiley.com/10.1029/2010WR009824`

Frensley, T., Crall, A., Stern, M., Jordan, R., Gray, S., Prysby, M., Newman, G., Hmelo-Silver, C., Mellor, D., & Huang, J. (2017). Bridging the Benefits of Online and Community Supported Citizen Science: A Case Study on Motivation and Retention with Conservation-Oriented Volunteers. *Citizen Science: Theory and Practice*, *2*(1), 4.
URL `https://theoryandpractice.citizenscienceassociation.org/article/10.5334/cstp.84/`

Gilbert, N. (2010). How to avert a global water crisis. *Nature*.
URL `http://www.nature.com/articles/news.2010.490`

Haklay, M. (2013). Citizen Science and Volunteered Geographic Information - overview and typoology of participation. In D. Sui, S. Elwood, & M. Goodchild (Eds.) *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, (pp. 105–122). Berlin: Springer.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., & Tallaksen, L. M. (2011). Large-scale river flow archives: importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200.
URL `http://doi.wiley.com/10.1002/hyp.7794`

Hobbs, S. J., & White, P. C. (2012). Motivations and barriers in relation to community participation in biodiversity recording. *Journal for Nature Conservation*, *20*(6), 364–373.
URL `http://dx.doi.org/10.1016/j.jnc.2012.08.002`

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., & Pierrefeu, G. (2018). Impact of Stage Measurement Errors on Streamflow Uncertainty. *Water Resources Research*, *54*(3), 1952–1976.
URL `http://doi.wiley.com/10.1002/2017WR022039`

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., & Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255.
URL `http://dx.doi.org/10.1080/02626667.2013.803183`

Johnson, M. F., Hannah, C., Acton, L., Popovici, R., Karanth, K. K., & Weinthal, E. (2014). Network environmentalism: Citizen scientists as agents for environmental advocacy. *Global Environmental Change*, *29*, 235–245.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0959378014001733`

Juston, J., Seibert, J., & Johansson, P.-o. (2009). Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, *23*(21), 3093–3109.
URL `http://doi.wiley.com/10.1002/hyp.7421`

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., Puntenney-Desmond, K., Seibert, J., & van Meerveld, I. (2018). Testing the waters: Mobile apps for crowdsourced streamflow data. *Eos*, *99*.

Kendall, M. G. (1990). *Rank Correlation Methods.*. London, UK & New York, NY: Oxford University Press, 5 ed.

Kirkwood, C. W. (1982). A Case History of Nuclear Power Plant Site Selection. *Journal of the Operational Research Society*, *33*(4), 353–363.
URL `https://www.tandfonline.com/doi/full/10.1057/jors.1982.77`

Kundzewicz, Z. W. (1997). Water resources for sustainable development. *Hydrological Sciences Journal*, *42*(4), 467–480.

Kundzewicz, Z. W. (2004). Editorial - Searching for change in hydrological data. *Hydrological Sciences Journal*, *49*(1), 3–6.
URL `https://www.tandfonline.com/doi/full/10.1623/hysj.49.1.3.53995`

Land-Zandstra, A. M., van Beusekom, M. M., Koppeschaar, C. E., & van den Broek, J. M. (2016). Motivation and learning impact of Dutch flu-trackers. *Journal of Science Communication*, *15*(1), 1–26.

Le Coz, J., Patalano, A., Collins, D., Guillén, N. F., García, C. M., Smart, G. M., Bind, J., Chiaverini, A., Boursicaud, R. L., Dramais, G., & Braud, I. (2016). Crowdsourced data for flood hydrology : feedback from recent citizen science projects in Argentina , France and New Zealand. *Journal of Hydrology*, *541*, 766–777.

Leeuw, T., & Boss, E. (2018). The HydroColor app: Above water measurements of remote sensing reflectance and turbidity using a smartphone camera. *Sensors*, *18*(256).

Levontin, L., Gilad, Z., & Chako, S. (2018). Questionare for the Motivation for Citizen Science Scale.
URL `https://www.cs-eu.net/news/questionare-motivation-citizen-science-scale`

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1-4), 272–288.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0022169497000413`

Loiselle, S., Thornhill, I., & Bailey, N. (2016). Citizen science: advantages of shallow versus deep participation. *Frontiers in Environmental Science*, *4*.
URL `http://www.frontiersin.org/10.3389/conf.FENVS.2016.01.00001/event{_}abstract`

Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, *51*(1), 151–156.
URL `http://doi.wiley.com/10.1111/j.1745-6584.2012.00956.x`

Lowry, C. S., Fienen, M. N., Hall, D. M., Stepenuck, K. F., & Paul, J. D. (2019). Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology. *frontiers in Earth Science*, *7*(128), 1–10.

Manning, R. (1891). On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, *20*, 161–207.

Matthias, S., Vitos, M., Altenbuchner, J., Conquest, G., Lewis, J., & Haklay, M. (2014). Taking Participatory Citizen Science to Extremes. *IEEE Pervasive Computing*, *13*(2), 20–29.
URL `http://ieeexplore.ieee.org/document/6818498/`

Meehan, T. D., Michel, N. L., & Rue, H. (2019). Spatial modeling of Audubon Christmas Bird Counts reveals fine-scale patterns and drivers of relative abundance trends. *Ecosphere*, *10*(4), e02707.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ecs2.2707`

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D., & Krysanova, V. (2015). On Critiques of "Stationarity is Dead: Whither Water Management?". *Water Resources Research*, *51*(9), 7785–7789.
URL `http://doi.wiley.com/10.1002/2015WR017408`

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290.
URL `https://linkinghub.elsevier.com/retrieve/pii/0022169470902556`

Njue, N., Stenfert Kroese, J., Gräf, J., Jacobs, S., Weeser, B., Breuer, L., & Rufino, M. (2019). Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects. *Science of The Total Environment*, *693*, 133531.
URL `https://linkinghub.elsevier.com/retrieve/pii/S0048969719334515`

Nov, O., Arazy, O., & Anderson, D. (2014). Scientists@Home: What drives the quantity and quality of online citizen science participation? *PLoS ONE*, *9*(4), 1–11.

Ogie, R., Clarke, R., Forehead, H., & Perez, P. (2019). Crowdsourced social media data for disaster management: Lessons from the PetaJakarta.org project. *Computers, Environment and Urban Systems*, *73*, 108–117.
URL `https://linkinghub.elsevier.com/retrieve/pii/S0198971518301066`

Peña, L., Murcia, H., Londoño, W., & Botina, H. (2017). Low-Cost Alternative for the Measurement of Water Levels in Surface Water Streams. *Sensors & Transducers Journal*, *217*(11), 36–44.
URL `http://www.sensorsportal.com/HTML/DIGEST/november{_}2017/Vol{_}217/P{_}2960.pdf`

Perrin, C., Ouding, L., Andreassian, V., Rojas-Serna, C., Michel, C., & Mathevet, T. (2007). Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, *52*(1), 131–151.
URL `http://www.tandfonline.com/doi/abs/10.1623/hysj.52.1.131`

Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, *554*, 613–622.

Raddick, J. M., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2013). Galaxy Zoo: Motivations of Citizen Scientists. *Astronomy Education Review*, *12*(1).
URL `http://www.portico.org/Portico/article?article=pgg3ztfcv7h`

Reges, H. W., Doesken, N., Turner, J., Newman, N., Bergantino, A., & Schwalbe, Z. (2016). CoCoRaHS: The Evolution and Accomplishments of a Volunteer Rain Gauge Network. *Bulletin of the American Meteorological Society*, *97*(10), 1831–1846.
URL `http://journals.ametsoc.org/doi/10.1175/BAMS-D-14-00213.1`

Rey-Mazón, P., Keysar, H., Dosemagen, S., D'Ignazio, C., & Blair, D. (2018). Public Lab: Community-Based Approaches to Urban and Environmental Health and Justice. *Science and Engineering Ethics*, *24*(3), 971–997.
URL `https://doi.org/10.1007/s11948-018-0059-8http://link.springer.com/10.1007/s11948-018-0059-8`

Rinderer, M., Kollegger, A., Fischer, B. M. C., Stähli, M., & Seibert, J. (2012). Sensing with boots and trousers - qualitative field observations of shallow soil moisture patterns. *Hydrological Processes*, *26*(26), 4112–4120.
URL `http://doi.wiley.com/10.1002/hyp.9531`

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0361476X99910202`

Sauermann, H., & Franzoni, C. (2015). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, *112*(3), 679–684.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, *103*(4), 663–688.
URL `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0029393`

See, L. (2019). A Review of Citizen Science and Crowdsourcing in Applications of Pluvial Flooding. *Frontiers in Earth Science*, *7*.
URL `https://www.frontiersin.org/article/10.3389/feart.2019.00044/full`

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224.
URL `http://www.hydrol-earth-syst-sci.net/4/215/2000/https://hal-sde.archives-ouvertes.fr/hal-00304596/`

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892.
URL `http://www.hydrol-earth-syst-sci.net/13/883/2009/`

Seibert, J., & McDonnell, J. J. (2015). Gauging the Ungauged Basin: Relative Value of Soft and Hard Data. *Journal of Hydrologic Engineering*, *20*(1), A4014004–1–6.
URL `https://ascelibrary.org/doi/10.1061/{%}28ASCE{%}29HE.1943-5584.0000861`

Seibert, J., van Meerveld, H., Etter, S., Strobl, B., Assendelft, R., & Hummer, P. (2019). Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen? *Hydrologie und Wasserbewirtschaftung*, *63*(2).

Seibert, J., & Vis, M. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325.
URL `http://www.hydrol-earth-syst-sci.net/16/3315/2012/`

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrological Processes*, *30*(14), 2498–2508.
URL `http://doi.wiley.com/10.1002/hyp.10887`

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125.
URL `http://doi.wiley.com/10.1002/hyp.11476`

Serrano Sanz, F., Holocher-Ertl, T., Kieslinger, B., Sanz Garcia, F., & Silva, C. G. (2014). White Paper on Citizen Science for Europe. Tech. rep., Socientize Consortium.
URL `https://ec.europa.eu/newsroom/dae/document.cfm?doc{_}id=6913`

Serret, H., Deguines, N., Jang, Y., Lois, G., & Julliard, R. (2019). Data Quality and Participant Engagement in Citizen Science: Comparing Two Approaches for Monitoring Pollinators in France and South Korea. *Citizen Science: Theory and Practice*, *4*(1), 22.
URL `https://theoryandpractice.citizenscienceassociation.org/article/10.5334/cstp.200/`

Seyfang, G., & Smith, A. (2007). Grassroots innovations for sustainable development: Towards a new research and policy agenda. *Environmental Politics*, *16*(4), 584–603.
URL `http://www.tandfonline.com/doi/full/10.1080/09644010701419121`

Sheppard, S. A., Turner, J., Thebault-Spieker, J., Zhu, H., & Terveen, L. (2017). Never Too Old, Cold or Dry to Watch the Sky. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–21.
URL `http://doi.acm.org/10.1145/3134729`

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, b. V., Krasny, M. E., & Bonney, R. (2012). Public Participation in Scientific Research: a Framework for Deliberate Design. *Ecology and Society*, *17*(2), art29.
URL `http://www.ecologyandsociety.org/vol17/iss2/art29/`

Sivapalan, M. (2003). Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170.
URL `http://doi.wiley.com/10.1002/hyp.5155`

Smith, L. C., Isacks, B. L., Bloom, A. L., & Murray, A. B. (1996). Estimation of Discharge From Three Braided Rivers Using Synthetic Aperture Radar Satellite Imagery: Potential Application to Ungaged Basins. *Water Resources Research*, *32*(7), 2021–2034.
URL `http://doi.wiley.com/10.1029/96WR00752`

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72.
URL `https://www.jstor.org/stable/1412159?origin=crossref`

Stepenuck, K. F., & Genskow, K. D. (2017). Characterizing the Breadth and Depth of Volunteer Water Monitoring Programs in the United States. *Environmental Management*, *61*(1), 46–57.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern-mental Panel on Climate Change. Tech. rep., Intergovernmental panel on climate change (IPCC), Cambridge, United Kingdom and New York, NY, USA.
URL `https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5{_}all{_}final.pdf`

Strasser, B. J., Baudry, J., Mahr, D., Sanchez, G., & Tancoigne, E. (2018). "Citizen Science"? Rethinking Science and Public Participation. *Science & Technology Studies*, *158887*, 52–76.
URL `https://sciencetechnologystudies.journal.fi/article/view/60425`

Strasser, B. J., & Haklay, M. (2018). Citizen Science : Expertise , Demokratie und öffentliche Partizipation. Tech. rep., Schweizerischer Wissenschaftsrat.

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019). The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. *PLOS ONE*, *14*(9), e0222579.
URL `http://dx.plos.org/10.1371/journal.pone.0222579`

Taguchi, T. (1939). Climatic change in historical time in Japan (2). *Journal of the Marine Meteorological Society*, *19*, 217–227 (in Japanese).

Thiel, S.-K., & Fröhlich, P. (2017). *Progress in Location-Based Services 2016*. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing.
URL `http://link.springer.com/10.1007/978-3-319-47289-8`

Thornhill, I., Loiselle, S., Clymans, W., & van Noordwijk, C. G. E. (2019). How citizen scientists can enrich freshwater science as contributors, collaborators, and co-creators. *Freshwater Science*, *38*(2), 231–235.
URL `https://www.journals.uchicago.edu/doi/10.1086/703378`

Uhlenbrook, S., Holocher, J., Leibundgut, C., & Seibert, J. (1998). Using a conceptual rainfall-runoff model on different scales by comparing a headwater with larger basins. *Water Resources and Ecology in Headwaters, IAHS Publication*, *248*, 297–305.

van Meerveld, H. J., Vis, M. J., & Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, *21*(9), 4895–4905.
URL `https://www.hydrol-earth-syst-sci.net/21/4895/2017/`

Venkatesh, Thong, & Xu (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157.
URL `http://www.jstor.org/stable/10.2307`

Walker, D., Forsythe, N., Parkin, G., & Gowing, J. (2016). Filling the observational void: Scientific value and quantitative validation of hydrometeorological data from a community-based monitoring programme. *Journal of Hydrology*, *538*, 713–725.
URL `https://linkinghub.elsevier.com/retrieve/pii/S0022169416302554`

Weeser, B., Jacobs, S., Kraft, P., Rufino, M. C., & Breuer, L. (2019). Rainfall - Runoff Modeling Using Crowdsourced Water Level Data. *Water Resources Research*, *55*, 1–16.

Weeser, B., Stenfert Kroese, J., Jacobs, S., Njue, N., Kemboi, Z., Ran, A., Rufino, M., & Breuer, L. (2018). Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of The Total Environment*, *631-632*, 1590–1599.
URL `https://linkinghub.elsevier.com/retrieve/pii/S0048969718308878`

West, S., & Pateman, R. (2016). Recruiting and Retaining Participants in Citizen Science: What Can Be Learned from the Volunteering Literature? *Citizen Science: Theory and Practice*, *1*(2), 1–10.
URL `http://theoryandpractice.citizenscienceassociation.org/articles/10.5334/cstp.8/`

World Water Assessment Programme (2003). *The 1st United Nations World Water Development Report: Water for People, Water for Life*. Paris: United Nations Educational, Scientific and Cultural Organization (UNESCO) and Berghahn Books.
URL `https://unesdoc.unesco.org/ark:/48223/pf0000129726`

Wright, D. R., Underhill, L. G., Keene, M., & Knight, A. T. (2015). Understanding the Motivations and Satisfactions of Volunteers to Improve the Effectiveness of Citizen Science Programs. *Society and Natural Resources*, *28*(9), 1013–1029.
URL `http://dx.doi.org/10.1080/08941920.2015.1054976`

# Paper I

frontiers in Earth Science

METHODS
published: 12 April 2019
doi: 10.3389/feart.2019.00070

# Virtual Staff Gauges for Crowd-Based Stream Level Observations

*Jan Seibert[1,2]*, Barbara Strobl[1], Simon Etter[1], Philipp Hummer[3] and H. J. (Ilja) van Meerveld[1]*

*[1] Department of Geography, University of Zurich, Zurich, Switzerland, [2] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3] SPOTTERON GmbH, Vienna, Austria*

Hydrological observations are crucial for decision making for a wide range of water resource challenges. Citizen science is a potentially useful approach to complement existing observation networks to obtain this data. Previous projects, such as CrowdHydrology, have demonstrated that it is possible to engage the public in contributing hydrological observations. However, hydrological citizen science projects related to streamflow have, so far, been based on the use of different kinds of instruments or installations; in the case of stream level observations, this is usually a staff gauge. While it may be relatively easy to install a staff gauge at a few river sites, the need for a physical installation makes it difficult to scale this type of citizen science approach to a larger number of sites because these gauges cannot be installed everywhere or by everyone. Here, we present a smartphone app that allows collection of stream level information at any place without any physical installation as an alternative approach. The approach is similar to geocaching, with the difference that instead of finding treasure-hunting sites, hydrological measurement sites can be generated by anyone and at any location and these sites can be found by the initiator or other citizen scientists to add another observation at another time. The app is based on a virtual staff gauge approach, where a picture of a staff gauge is digitally inserted into a photo of a stream bank or a bridge pillar, and the stream level during a subsequent field visit to that site is compared to the staff gauge on the first picture. The first experiences with the use of the app by citizen scientists were largely encouraging but also highlight a few challenges and possible improvements.

**Keywords: citizen science, smartphone app, water level class, crowdsourcing, data collection**



***Edited by:***
*Jonathan D. Paul,*
*Imperial College London,*
*United Kingdom*

***Reviewed by:***
*Wouter Buytaert,*
*Imperial College London,*
*United Kingdom*
*Tim van Emmerik,*
*Delft University of Technology,*
*Netherlands*
*Jon Olav Skøien,*
*European Commission – Joint*
*Research Center, Belgium*

****Correspondence:***
*Jan Seibert*
*jan.seibert@geo.uzh.ch*

***Specialty section:***
*This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science*

***Received:*** *07 September 2018*
***Accepted:*** *19 March 2019*
***Published:*** *12 April 2019*

***Citation:***
*Seibert J, Strobl B, Etter S, Hummer P and van Meerveld HJ (2019) Virtual Staff Gauges for Crowd-Based Stream Level Observations. Front. Earth Sci. 7:70. doi: 10.3389/feart.2019.00070*

## INTRODUCTION

Data on the quantity and quality of water are needed for appropriate water management decisions. However, hydrology and water resources management are frequently restricted by limited data availability, particularly in data-scarce regions with urgent water management issues (Mulligan, 2013). The decline of national hydrological and meteorological observation networks (Vörösmarty et al., 2001; Fekete et al., 2012; Ruhi et al., 2018) is frustrating, especially in light of the current local and global water-related challenges, and those ahead, such as adaptation to extreme events and securing water resources for a growing population. Although new observation techniques, including remote sensing, geophysical methods, and wireless sensor networks, provide exciting opportunities for new data collection, central hydrological variables, such as soil moisture or streamflow remain difficult to observe with a sufficient spatiotemporal resolution. Therefore, crowd-based data collection might be a valuable complementary approach to collect data and overcome data limitations (Buytaert et al., 2014).

The idea to include the public in hydrological and meteorological data collection is by no means new. The Swedish meteorologist Tor Bergeron asked the public through appeals over radio and phone calls to measure snow depth (Bergeron, 1949) and rainfall (Bergeron, 1960) and to mail their observations on postcards. This resulted in much more detailed maps than would have been possible with official station data alone. It allowed the creation of a snow depth map for an area of one degree square covering Uppland, Sweden based on 98 observations by volunteers rather than data from only 12 official stations (Bergeron, 1949). For the rainfall observations, Bergeron and his co-workers developed the Pluvius rain gauge as an inexpensive alternative to existing, official gauges. While later there were ~800 of these gauges in other parts in Sweden, for the initial surveys during 1953 about 150 gauges were distributed in a ~30 km by ~30 km area around Uppsala, Sweden (Bergeron, 1960). Both of these projects led to a better understanding of the influence of topography and vegetation on precipitation formation. Even though these early studies were very successful, similar approaches remained rare due to the logistical challenge to transmit and enter the collected data in a common database. However, recent developments in information and communication technology provide exciting new opportunities for citizen-science based approaches using text messages (Lowry and Fienen, 2013; Weeser et al., 2018), websites (e.g., Stream Tracker[1]), apps (e.g., Teacher et al., 2013; Davids et al., 2018; Kampf et al., 2018; Photrack[2]), data mining (Smith et al., 2015; Li et al., 2018) or custom-designed wearable sensors (e.g., Hut et al., 2016; smartfin[3]). However, as stated by Jerad Bales, the Chief scientist for hydrology at the U.S. Geological Survey, "*Crowdsourcing water-information is in its infancy [...], and there remain major issues of data quality and sustainability (Lowry and Fienen, 2013). Nevertheless, the use of crowdsourcing to report routine water data, as well as information on floods and droughts, needs to be creatively explored*" (Bales, 2014).

With a large number of contributions from citizens, the CrowdHydrology project[4] (Lowry and Fienen, 2013) has (and still does) successfully demonstrated that it is possible to engage the public in hydrological measurements by asking them to submit stream level observations via text messages. A similar system was implemented in Cithyd[5]. However, these approaches using staff gauges (scaled measurement sticks in the water) restrict the number of places where stream levels can be observed because staff gauges cannot be installed everywhere and by everyone. In mountainous streams, a stable installation is challenging even for hydrologists, and often permits are required before a staff gauge can be installed. Furthermore, if a physical installation is possible, one might consider installing a stream level logger instead of a staff gauge as these loggers have become less expensive and more reliable in recent years. Instead, we propose an approach where anyone can start a measurement location and the observations can be taken anywhere and by anyone. Our approach is similar to geocaching[6], with the difference that instead of treasure hunting sites, stream level observation sites are established and can be revisited by other citizen scientists. In this paper, we describe the virtual staff gauge approach, highlight several design considerations, and discuss whether people understand the concept. In another study (Strobl et al., 2019), we found that most people can classify the water level correctly by comparing it to a reference picture with a virtual staff gauge. Here the focus was on how well people are able to "install" a virtual staff gauge in the app, i.e., taking the reference picture and placing the staff gauge in this picture.

[1] http://www.streamtracker.org
[2] http://www.photrack.ch/mobile.html
[3] https://smartfin.org/
[4] http://www.crowdhydrology.com
[5] http://www.cithyd.com/it/

## VIRTUAL STAFF GAUGE

### General Approach

The advantage of the virtual staff gauge approach is that it avoids physical installations and makes the setup of new observation sites fast and easy. The basic idea behind our approach for stream level observations is that it is usually possible to identify a number of features in a stream or on the streambank, such as rocks, that allow ranking of the stream levels (i.e., "below this tree but above that rock"). While such stream level class observations are not as precise as continuous stream level observations from a staff gauge (i.e., no millimeter resolution) and provide more qualitative information such as "the water level is very low" or "there is a flood event," they can be quite informative for hydrological modeling (van Meerveld et al., 2017). The challenge is to allow easy identification of the different stream level classes, without the need for lengthy verbal descriptions. A picture is helpful in this respect but needs to be amended by a scale. For this, we use the virtual staff gauge approach (see also **Figure 1**):

- The user chooses a suitable site along a stream and identifies the location on a map in the smartphone app.
- The user takes a picture of the streambank (perpendicular to the flow direction and as level as possible, to minimize contortion of the view). There should be some reference in the picture, such as a bridge or stones and ideally, the picture is taken during low flow conditions.
- An image of a yardstick with a number of classes is digitally inserted into the picture as a virtual staff gauge. The user can move the inserted staff gauge in the image and scale it so that it covers the expected stream level variations.

[6] https://www.geocaching.com/

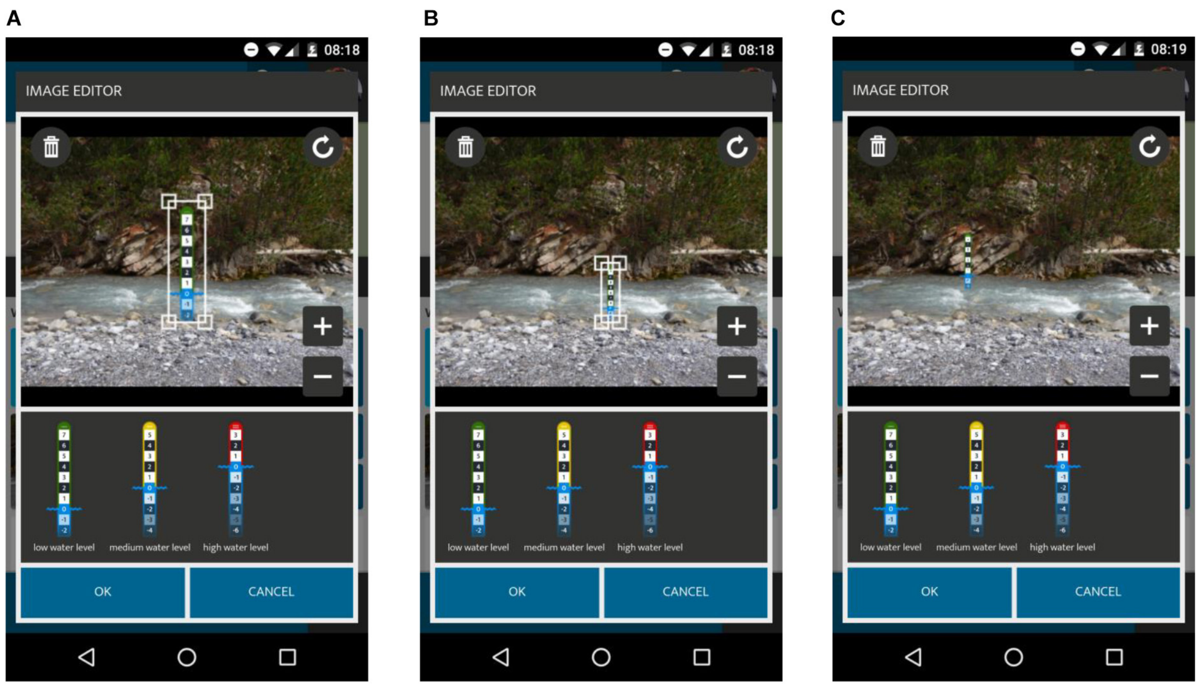



**FIGURE 1 |** Series of screenshots showing the insertion of the virtual staff gauge in the reference picture: **(A)** insert the image of the staff gauge in the reference picture, **(B)** scale the inserted image, and **(C)** move the image so that the blue line matches the stream level in the picture.

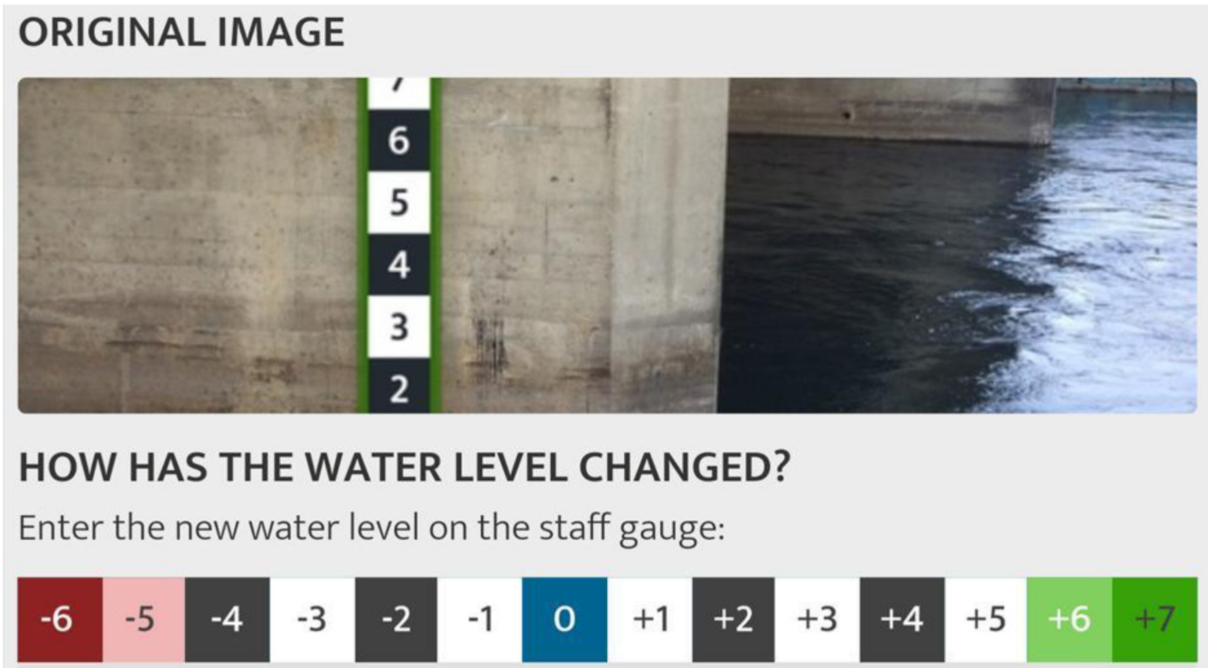



**FIGURE 2 |** The horizontal version of the staff gauge at the "Update Spot" interface as selectable buttons to report the new water level class observation. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

This reference picture with the virtual staff gauge allows anyone who visits the site at a later time to estimate the stream level class by relating the current stream level to the features on the photo and the virtual staff gauge (e.g., the stream level has changed and is now above a certain rock). For this update, a simplified horizontal staff gauge design is used in the "Update

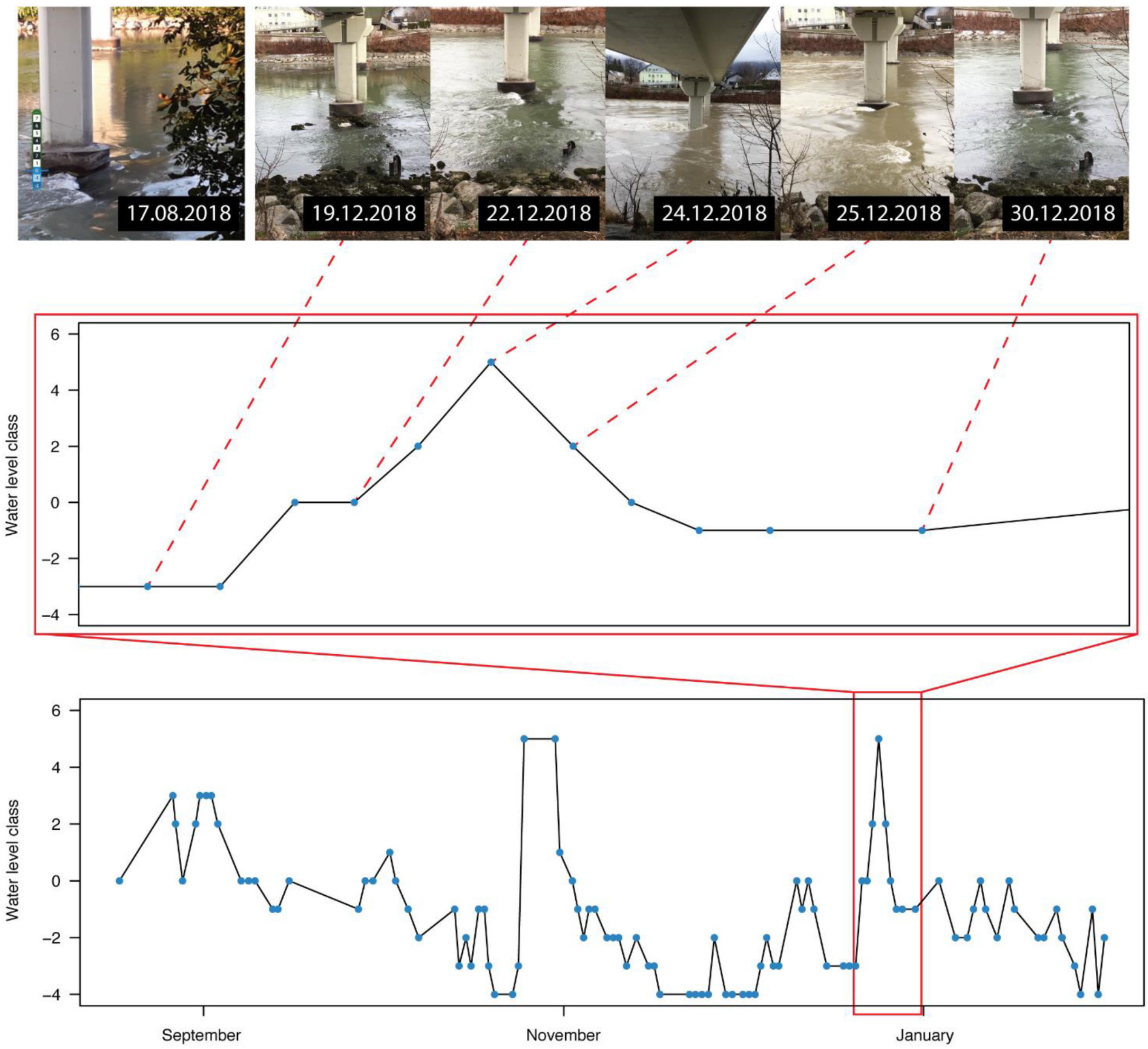



**FIGURE 3 |** Example of a water level time series obtained using the CrowdWater app (River Salzach, Austria). The pictures for one runoff event (and the reference picture) are shown as an example in the top row.

Spot" interface of the app (**Figure 2**) that shows the full range of class bars for input. To update a spot and provide a new observation of the stream level, the user compares the current stream level with the reference picture with the staff gauge in the app, takes a new picture of the stream, clicks on the current stream level class on the horizontal staff gauge and submits the new observation to the data servers. Over time, this results in a time series of water level observations (**Figure 3**). It is important to note, that the user observes and enters the water level; the new picture is only used for documentation. While automated image recognition could be valuable, at this point we rather rely on human eyes and interpretation and avoid issues such as the exact location and angle when the picture is taken. The pictures, however, allow data quality control. We have recently developed the CrowdWater game as an approach to use these pictures for crowdbased quality control of the water level class data (see "Game"[7]).

## Design Considerations and Initial Tests

Several decisions on the design of the virtual staff gauge had to be taken before implementation in the smartphone app. Early on it was decided to use relative stream level classes instead of numeric values in, for instance, centimeters, as there is an obvious limitation in the resolution of stream-level observations that can be achieved with a virtual staff gauge. Translating the virtual staff gauge levels to absolute levels would also make the "virtual installation" much more time consuming as it would require observations of different heights.

[7] https://www.crowdwater.ch

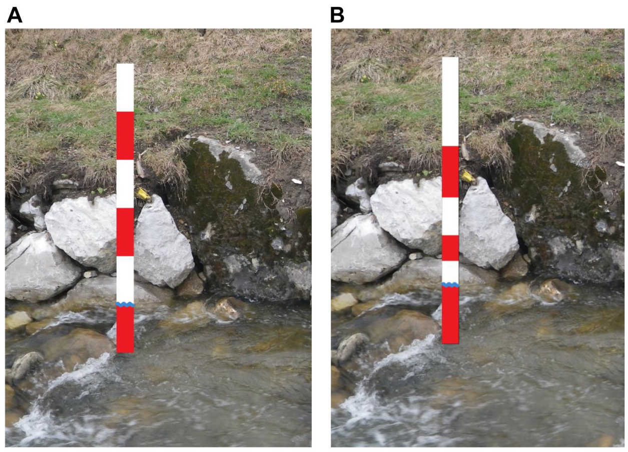



**FIGURE 4 |** Early version of the virtual staff gauge with regular **(A)** and irregular **(B)** class sizes.

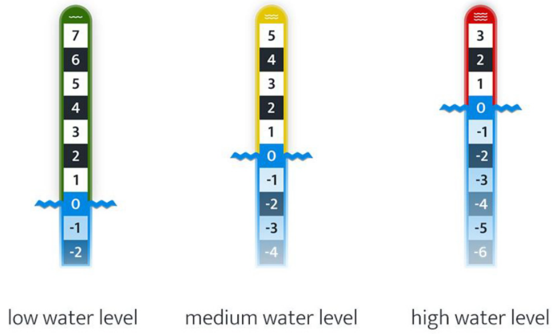



**FIGURE 6 |** The three staff gauges available in the app. Their ideal application depends on the flow condition of the river at the time that the reference picture is taken. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

Absolute levels would also be site-specific, i.e., the offset would vary largely from place to place. Fortunately, absolute levels are not needed for the potential use in hydrological modeling because the relative values provide important information on the timing of streamflow responses (Seibert and Vis, 2016; van Meerveld et al., 2017).

In an early test with university students, two different types of staff gauges were tested. In addition to regular class sizes (as ultimately implemented in the app), we also tested irregular class sizes (**Figure 4**), but this idea was discarded because some users found it confusing and because it did not allow for as much flexibility as we had hoped.

Once we had decided to have a non-metric virtual staff gauge with regular class sizes, we started to discuss the implementation with SPOTTERON, which is the app company hired to develop the CrowdWater app. During these discussions, the focus was largely on how to make the app intuitive to use. A clearly visible blue wave on the virtual staff gauge was chosen to indicate the stream level at the time that the reference picture was taken (**Figure 5**). During placement, the citizen scientists will highlight the stream level in the photo with the water line in the staff gauge (**Figure 1**). We decided to use ten classes on the virtual staff gauge; this was a compromise between simplicity, resolution, and usability. Through the use of a negative and positive scale, we tried to make the image even more intuitive, as a negative value

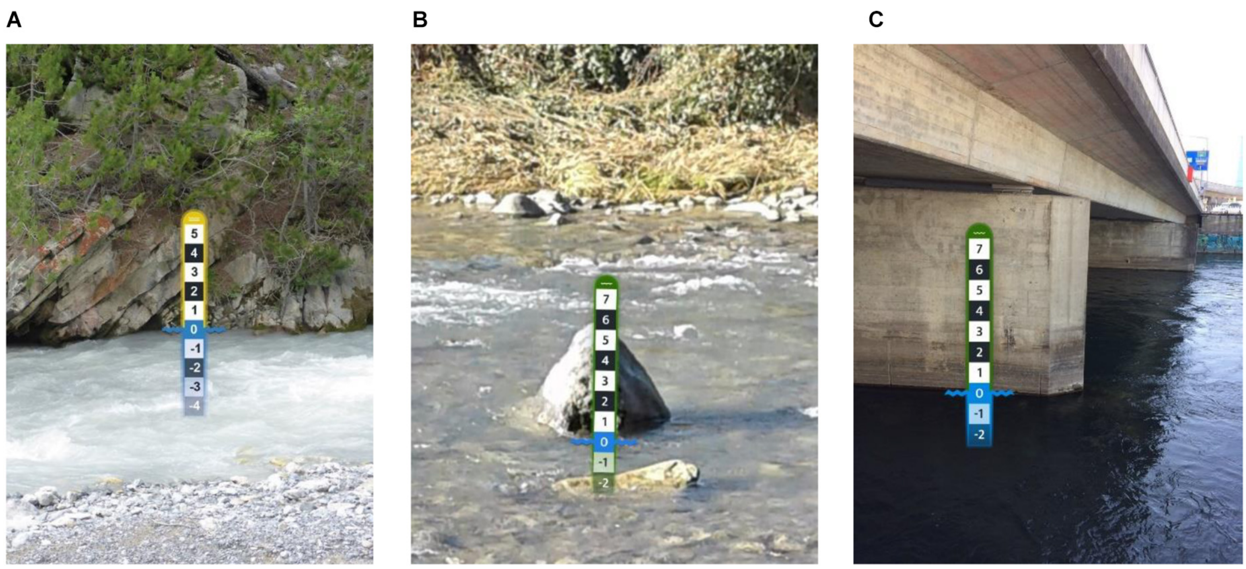



**FIGURE 5 |** Examples of well-placed virtual staff gauges on **(A)** the opposite stream bank, **(B)** a rock in the stream, and **(C)** a bridge pillar, showing the blue wave at the stream level when the site was established and the positive and negative scale above and below the current stream level, respectively.

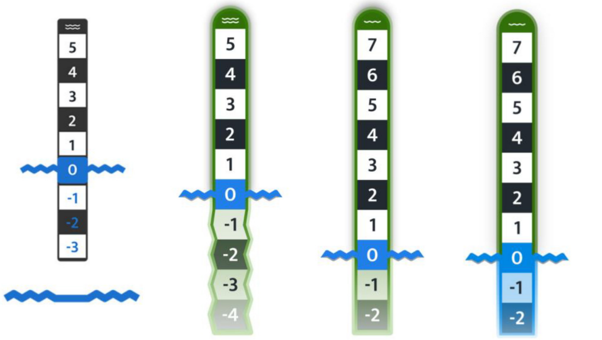



**FIGURE 7 |** Various staff gauge designs. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

would indicate a stream level below the level in the reference picture and a positive value above it (**Figure 6**). The stream level numbers and class bars follow a neutral black/white scheme to utilize contrast between the sections but also maintain secondary visual weight.

We recommend that citizen scientists initiate a new measurement site during low flow conditions because the reference points are better visible during low flow conditions and this enables future users to better assess the situation for an update. However, this might be a strong restriction in practice and we, therefore, decided to allow insertion of virtual staff gauges also in photos taken during situations with high stream levels. To use suitable staff gauges for all flow conditions, we decided to offer three different staff gauges to the user (**Figure 6**). The green staff gauge is best suited for rivers with a low water level at the time that the reference picture is taken, as it still has many positive classes (i.e., above the blue wave) to record stream levels for higher flow conditions. The yellow staff gauge is well suited for when the reference picture is taken at average flow conditions, and the red staff gauge is ideal for high flow conditions. The red, yellow and green staff gauges were chosen because strong, vibrant colors visually communicate not only a difference but also a development over time, e.g., traffic lights signal different states of movement.

## Virtual Staff Gauge Implementation

The virtual staff gauge was implemented as a so-called "sticker". Stickers are a common practice in app design; they use image- or vector-based content as overlays in photos that are taken on a smartphone. They are mainly used in messenger tools, such as WhatsApp or Facebook Messenger to add additional information or emotions to images. Positioning and transformation are usually done by multi-touch gestures for scaling, placement, and rotation. In this case the sticker has to be moved so that the staff gauge is aligned with the streambank or bridge pillar and the blue line is located at the water level (**Figure 1**). By adopting such a rather well-known input method, the use of the app is more intuitive and, thus, optimizes usability. Obviously, using an established technique also had technical advantages for the implementation.

In practice, the placement of the staff gauge can happen on bright or dark, blurry or clear, high- or low-saturation pictures, taken by the users on all kinds of smartphone models and cameras. Therefore, various designs for the virtual staff gauges were tested on different backdrop images and directly on smartphone screens (**Figures 7**, **8**). To ensure that the staff gauge is visible in various conditions, we used additional soft shadows to enhance the edge contrast, but still let the staff gauge immerse itself into the picture as part of the scenery. We furthermore decided to strengthen the visual representation of the areas above and below the stream level by using a blue hue for all class bars below the water level and making them slightly transparent (**Figures 6–8**).

# TEST OF THE APP IN PRACTICE

## CrowdWater App

The virtual staff gauge was implemented in the CrowdWater smartphone app. The app was first launched for iOS and

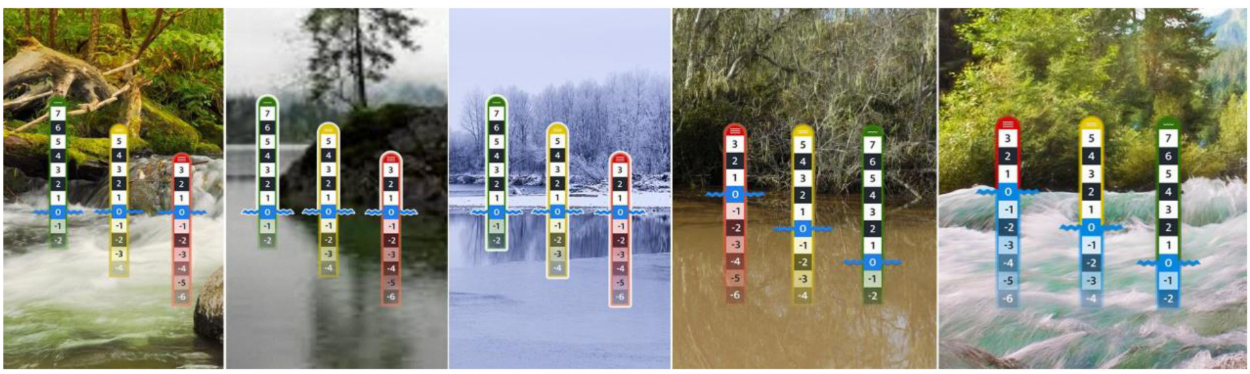


**FIGURE 8 |** Staff gauge design variants in different environments. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net. Note that the virtual staff gauges were not scaled nor placed correctly (see **Figure 1**).

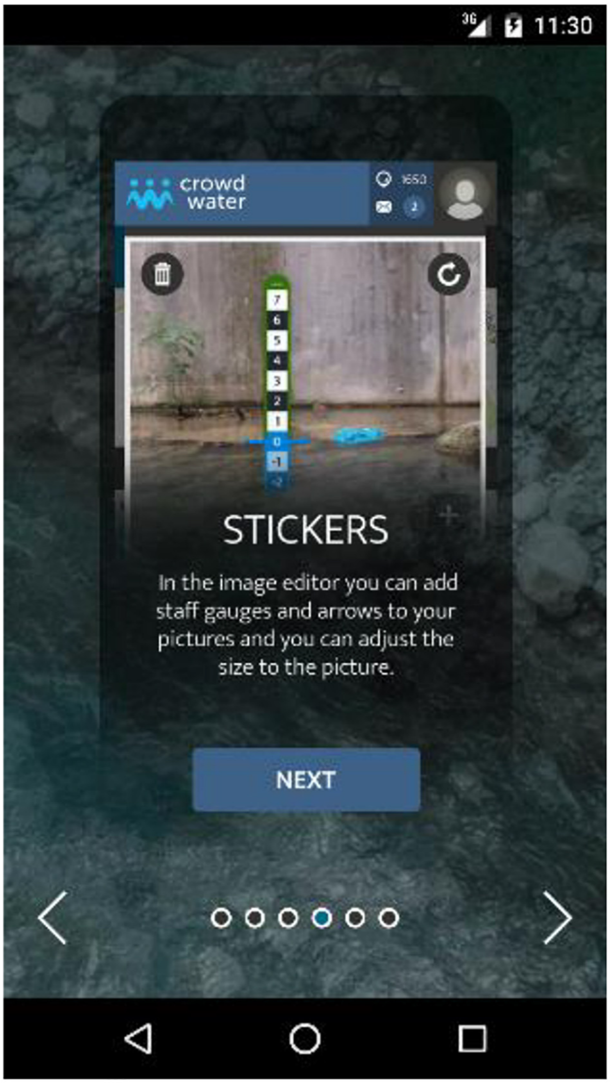



**FIGURE 9 |** Screenshot of an intro slide that appears when the app is opened for the first time. These can be re-watched anytime. The goal is to quickly provide the most important information on the basic functionalities of the app.

Android in March 2017; there have been several updates of the app since its initial launch. The app was promoted on the CrowdWater homepage (see Footnote 7), through Facebook, Twitter, Instagram, LinkedIn, and ResearchGate posts, as well as on the CrowdWater YouTube channel and at several conferences.

When starting the app, the user has to browse through a number of intro-slides that explain the basic functionalities and the interface of the app. Among them is the sticker function of the virtual staff gauge (**Figure 9**). Additional guidance on how to use the app in the form of texts, pictures and videos are provided on the project homepage and in an explanatory YouTube video[8].

[8]https://www.youtube.com/watch?v=3ag4sHWf0yg

**TABLE 1 |** Collection of errors made by app-users grouped into broader error categories and frequency of occurrence.

| Error type | | Frequency of occurrence |
|---|---|---|
| Staff gauge size problem | Staff gauge too big | +++ |
| | Staff gauge too small | + |
| Staff gauge placement problem | Wrong angle | +++ |
| | Staff gauge not on the water surface | +++ |
| Unsuitable location | Lack of reference structure for stream level identification | ++ |
| | Structure hidden by vegetation or snow | + |
| | Unclear which structure to use | + |
| | River bank too far away | ++ |
| | Poor image quality | + |
| | Site not easily accessible | . |
| | No suitable site for staff gauge placement available | . |
| | Changes in the rating curve | + |
| | Multiple measurement sites at (almost) the same location | + |
| | Testing (e.g., beer glasses, not a river, out of a train, etc.) | ++ |

*+++: occasional = more than 10 times; ++: seldom = 5–10 times; +: rare: less than 5 times; . : not quantifiable.*

## Typical Mistakes

While users seem to understand the approach used in the CrowdWater app in general, there were also a number of recurrent mistakes related to the staff gauge placement or size. These mistakes affect about 10% of the more than 500 reference pictures (**Table 1**). Staff gauge placement or size problems could be due to users not having read the available instruction material or not fully understanding the concept. Some other issues are not directly related to setting up a virtual staff gauge site but still affect the results, e.g., it is less useful if users create new measurement sites in, or close to, a location where another spot already exists than when they update the existing spot or start a new site on a different river.

### Staff Gauge Placement Problem

The most common mistake was related to the placement of the virtual staff gauge. Some users took pictures in the direction of the flow (instead of perpendicular to the flow, see example in **Figure 10**). This makes it almost impossible to place a virtual staff gauge that allows subsequent level observations because clear reference features are usually missing on these pictures. Another placement related issue occurs when the blue wave of the staff gauge is not located at the water surface in the reference picture. This means that the stream level of the reference picture is not at zero, which could lead to confusion for other users when updating the spot later on.

### Staff Gauge Size Problems

In a number of cases, the size of the staff gauge was suboptimal. This may be either because people do not realize that they

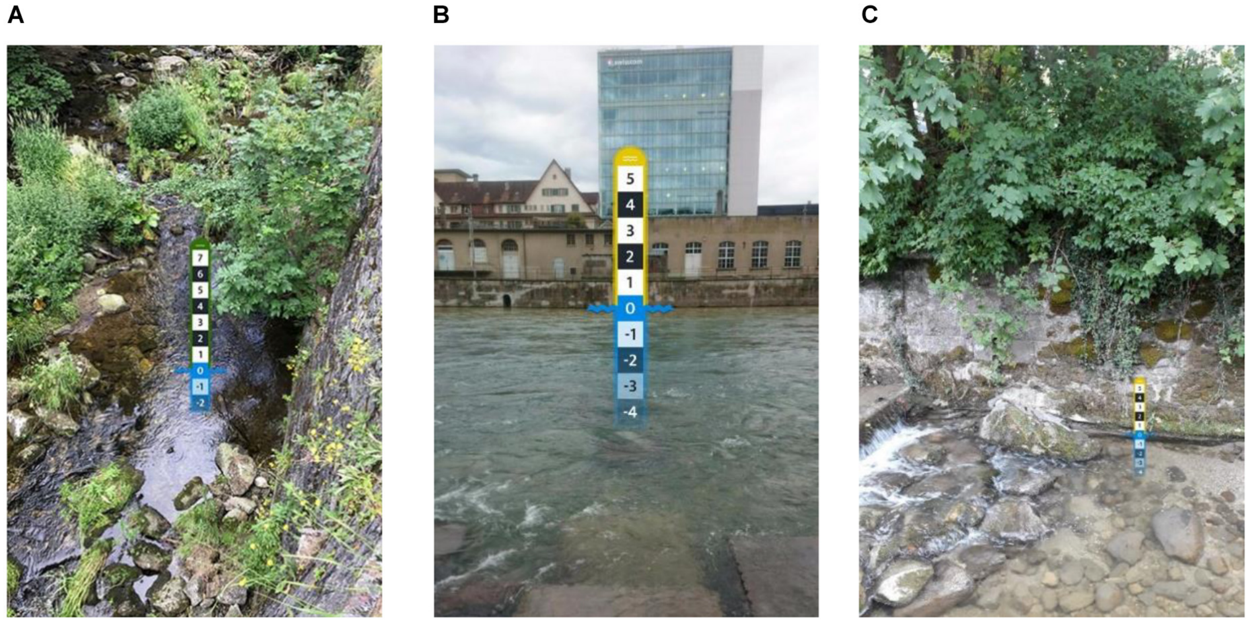



**FIGURE 10 |** Examples of misplaced virtual staff gauges: **(A)** The picture was taken in the upstream direction instead of perpendicular to the flow direction, which makes it impossible to estimate subsequent stream level changes, **(B)** The virtual staff gauge is so large that it is unlikely that the water level will reach different classes and is therefore improbable to obtain an approximate representation of the stream hydrograph, **(C)** The small virtual staff gauge can show small changes in the stream level, but cannot represent very high flows as anything above a medium flow falls into the highest class.

can resize the size of the staff gauge or do not understand why it is useful to rescale the staff gauge. The perfect staff gauge size is however, somewhat subjective and might to some degree depend on the specific research question and data needs for a site.

In our instruction material, we show the optimal case where the highest class of the staff gauge reaches up to the level of the highest in-bank flow. This may, however, be hard to imagine for citizen scientists and is probably also not considered when users place their first virtual staff gauge. Staff gauges that are too large are not only unrealistic (i.e., the stream level is very unlikely to rise into the highest classes) but this also reduces the variation in future observations because it is less likely that a change in stream level is large enough to reach the next class. There were also a few cases where the staff gauge was too small. A small staff gauge can make it hard to determine the class of the current stream level because the differences between the classes are too small. It also makes it hard to document very high or very low flows. Furthermore, finding the location of the measurement site can be challenging when users take a very detailed (zoomed-in) picture of the reference structure. This issue was more common for small staff gauges and could probably be solved by implementing an option to add an overview photo that shows the general location of the reference structure.

## Unsuitable Location

An obvious problem are pictures that lack references for level identification or pictures where a staff gauge was not inserted in the picture. Optimal conditions to place a virtual staff gauge, such as a vertical wall on the opposite river bank or a vertical structure like a rock or bridge pillar in the river, are sometimes hard to find. At least in some cases, the reason for problematic pictures could also be that the rivers were not easily accessible or had no suitable reference features but people still wanted to take a picture to establish a measurement site. Another problem is that in some locations the vegetation growth obscures features on the river bank that were visible when the reference picture was taken (e.g., in winter when there was no vegetation). This makes it nearly impossible to compare stream levels properly. Reference pictures with snow can also make it difficult to assess the stream level later on.

On wide rivers, it is difficult to place a reasonably sized staff gauge at the opposite river bank and still observe changes in stream levels. Furthermore, in these cases, the quality of the pictures is often low due to zooming. This problem can be solved at locations with an instream structure (such as a bridge pillar) and placing the staff gauge along a pillar.

Changes due to erosion or sedimentation are another issue. In these cases stream levels are not a reliable indicator of streamflow. Our dataset contains one site where the riverbed changed quite drastically due to deposited sediment. Because the reference structure (a concrete wall next to a bridge) stayed in place, approximately the same flow meant a different stream level class compared to the situation in the reference picture taken before the sediment was deposited. The solution to this problem would be to archive the reference picture and create a new one.

## CONCLUDING REMARKS

In this paper, we presented a new citizen science approach based on virtual staff gauges that allow crowd-based stream level observations along any stream. The advantage of this approach is that no physical installations are needed, which makes the approach fully scalable, as it is easy and quick for anyone to set up a new measurement site or contribute an observation to an existing site. As discussed in this paper, during development and testing of the virtual staff gauge approach, we identified several issues that required modifications in the original design. Further app developments and better guidance for app users on how to set up a virtual staff gauge site will reduce the number of incorrect sites in the future. Despite these challenges, the first experiences from using the virtual staff gauge approach are encouraging and show that this approach can be useful to collect stream level data at many locations by citizen scientists.

In the first year since launching the smartphone app, numerous measurement sites have been set up. On 3. September 2018, 2431 observations had been submitted by 218 users. For 79 of the 675 sites, more than five updates on the stream level class had been submitted. The collected data have a limited resolution due to the use of stream level classes and are sometimes spotty in time. However, previous work using synthetic data indicates that such data are still informative to constrain hydrological models. Time series of precipitation and temperature are more likely to be available than those of streamflow. The observed stream level class data can, thus, be used in combination with these time series to generate modeled streamflow time series. The potential value of such data has been evaluated based on subsets of existing data. These studies have indicated the value of water level class data for model calibration (van Meerveld et al., 2017); uncertain streamflow estimates were less informative (Etter et al., 2018). The water level data collected in the CrowdWater project are publicly available, and we expect them also to be used for other uses, be it for research, flood protection or leisure activities.

While our current focus is on measurement sites in Switzerland, the app can be, and is already, used worldwide. For developing and evaluating the value of the data obtained with the virtual staff gauge approach countries with a relative wealth of stream data, such as Switzerland, are favorable, but we anticipate that, once developed and tested, the approach will be most beneficial in regions where data are scarce.

## AUTHOR CONTRIBUTIONS

JS and HvM developed the first idea of the virtual staff gauge while hiking along a Swiss creek. BS and SE were responsible for the tests and the evaluation of the user experience of the app and contributed by specifying the requirements for the app, which were then discussed among all authors and further developed with PH. PH was responsible for most of the graphical design and the implementation of the smartphone app. JS wrote the manuscript with input from all authors.

## FUNDING

The CrowdWater project is funded by the Swiss National Science Foundation (Project Number 163008).

## ACKNOWLEDGMENTS

We thank all participants of the CrowdWater project for contributing their observations.

## REFERENCES

Bales, J. D. (2014). Progress in data collection and dissemination in water resources - 1974-2014. *Water Resour. Impact* 16, 18–23.

Bergeron, T. (1949). The problem of artificial control of rainfall on the globe. Part II: the coastal orographic maxima of precipitation in autumn and winter. *Tellus* 1, 15–32. doi: 10.1111/j.2153-3490.1949.tb01264.x

Bergeron, T. (1960). *Operation and Results of "Project Pluvius"*. Washington, DC: American Geophysical Union, 152–157. doi: 10.1029/GM005p0152

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., et al. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Front. Earth Sci.* 2:26. doi: 10.3389/feart.2014.00026

Davids, J. C., Rutten, M. M., Shah, R. D. T., Shah, D. N., Devkota, N., Izeboud, P., et al. (2018). Quantifying the connections—linkages between land-use and water in the Kathmandu Valley. *Nepal. Environ. Monit. Assess.* 190:17. doi: 10.1007/s10661-018-6687-2

Etter, S., Strobl, B., Seibert, J., and van Meerveld, I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrol. Earth Syst. Sci.* 22, 5243–5257. doi: 10.5194/hess-2018-355

Fekete, B. M., Looser, U., Pietroniro, A., and Robarts, R. D. (2012). Rationale for monitoring discharge on the ground. *J. Hydrometeorol.* 13, 1977–1986. doi: 10.1175/JHM-D-11-0126.1

Hut, R., Tyler, S., and Van Emmerik, T. (2016). Proof of concept: temperature-sensing waders for environmental sciences. *Geosci. Instrum. Methods Data Syst.* 5, 45–51. doi: 10.5194/gi-5-45-2016

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., et al. (2018). Testing the waters: mobile apps for crowdsourced streamflow data. *EOS* 99, 30–34. doi: 10.1029/2018EO096355

Li, Z., Wang, C., Emrich, C. T., and Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartogr. Geogr. Inf. Sci.* 45, 97–110. doi: 10.1080/15230406.2016.1271356

Lowry, C. S., and Fienen, M. N. (2013). CrowdHydrology: crowdsourcing hydrologic data and engaging citizen scientists. *Ground Water* 51, 151–156. doi: 10.1111/j.1745-6584.2012.00956.x

Mulligan, M. (2013). WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrol. Res.* 44:748. doi: 10.2166/nh.2012.217

Ruhi, A., Messager, M. L., and Olden, J. D. (2018). Tracking the pulse of the Earth's fresh waters. *Nat. Sustain.* 1, 198–203. doi: 10.1038/s41893-018-0047-7

Seibert, J., and Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrol. Process.* 30, 2498–2508. doi: 10.1002/hyp.10887

Smith, L., Liang, Q., James, P., and Lin, W. (2015). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *J. Flood Risk Manag.* 10, 370–380. doi: 10.1111/jfr3.12154

Strobl, B., Etter, S., van Meerveld, I., and Seibert, J. (2019). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrol. Sci. J.* (in press). doi: 10.1080/02626667.2019.1578966

Teacher, A. G. F., Griffiths, D. J., Hodgson, D. J., and Inger, R. (2013). Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecol. Evol.* 3, 5268–5278. doi: 10.1002/ece3.888

van Meerveld, H. J., Vis, M. J. P., and Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrol. Earth Syst. Sci.* 21, 4895–4905. doi: 10.5194/hess-21-4895-2017

Vörösmarty, C. J., Askew, A., Grabs, W., Barry, R. G., Birkett, C., Döll, P., et al. (2001). Global water data: a newly endangered species. *Eos Trans. Am. Geophys. Union* 82, 1999–2001. doi: 10.1029/01EO00031

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., et al. (2018). Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Sci. Total Environ.* 631–632, 1590–1599. doi: 10.1016/j.scitotenv.2018.03.130

**Conflict of Interest Statement:** PH is founder and co-owner of the company SPOTTERON GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Paper II

# Why Do People Participate in Environment-Focused Citizen Science Projects?

Simon Etter, Barbara Strobl, Jan Seibert, H. J. (Ilja) van Meerveld, Kai Niebert

## Key Findings:

- The motivation of participants in two environment-focused citizen science projects was evaluated using an online questionnaire. The results were classified using two categorizations of motivations for citizen science projects from the literature.
- An interest in science and the project's topic were the main motivations to join for participants of both projects.
- Participants of CrowdWater were more motivated by conformity than participants of Naturkalender.
- For participants of Naturkalender the activities matched previous experiences. They wanted to share their knowledge and experience and more frequently highlighted the fun aspect for their initial participation than participants of the CrowdWater project.
- Participants in the 50-59 age group were most motivated by breaking their everyday routine, being outside, learning something new and challenging themselves. Participants in the other age groups were most motivated by contributing to science.
- Feedback to participants can be provided by project administrators and also by other users through social media elements.

# Abstract

We investigated the motivations of participants in two environment-focused citizen science projects using an online questionnaire. The two projects, CrowdWater and Naturkalender (English: Nature's Calendar), aim to collect data on water and phenology, respectively, and use similar smartphone apps. Our questions focused on both the motivations for initial engagement and in how far these are fulfilled by participating in the citizen science projects. For the questionnaire, we used a set of statements based on responses to open questions from the citizen science and volunteering literature. The questionnaire was sent to all participants of the projects. The answers were analysed based on two different categorisation schemes. We found that the motivations to participate in the projects were similar for the two projects but there were also some differences. The main motivations for becoming engaged in the projects were to contribute to science, to improve the wellbeing of society and to protect nature. The CrowdWater participants were in general more motivated by conformity (i.e., being asked to participate or social pressure) than the Naturkalender participants. Participants of the Naturkalender project and participants in the 50-59-year age group of both projects agreed most to enjoying their participation and learning something new. Super-users, i.e., users who participate at least once per week, were motivated more by contributing to science and the competitive elements of the projects than the occasional participants. Many of the participants who joined because they were asked directly and felt obliged to do so, submitted only a few observations. Based on the results of this study and previous studies reported in the literature, we recommend that to improve engagement and retention of participants, projects should aim to find people who are already interested in the topic, have a related hobby, or are affected by the problem that the project tries to solve. Furthermore, it is beneficial if feedback to participants is provided by project administrators or other participants (e.g. using social media elements).

# 1 Introduction

The number of citizen science projects is growing rapidly (Irwin, 2018). For all projects it is important to lower the hurdles for sustained participation (Domroese and Johnson, 2017). This includes designing projects that meet people's interest and communicating with the participants. Engagement in citizen science projects depends strongly on motivational factors (Phillips et al., 2019). Understanding these factors is, thus, important for project managers. However, the motivations of people to participate in citizen science and how people benefit from participation are complex and require more research (Haklay, 2018; Thornhill et al., 2019; West and Pateman, 2016). The attitudinal construct of motivation has been used in different contexts, such as learning of students (Martin, 2007) and volunteerism (Bell et al., 2008). Phillips et al. (2018) define motivation as "a form of goal setting to achieve a behaviour or result" but also state that the term motivation has not been used consistently in the field of citizen science. They, furthermore, argue that many studies that claim to report citizen's motivations, actually report reasons to participate (e.g., the desire to help science) instead of the psychological underpinnings of behaviour (e.g., "because it makes me feel good"). For this manuscript, we adopt the definition of Phillips et al. (2018), which also includes reasons to participate. We consider motivations and reasons equally important for the successful management of citizen science projects. This is in line with other studies on motivations or reasons of citizen scientists (Hobbs and White, 2012; Raddick et al., 2010).

The main motivations to join citizen science projects, reported so far, are to contribute to science and to protect the environment, as well as to be part of a specific community (Alender, 2016; Curtis, 2015; Raddick et al., 2013). Johnson et al. (2014) used open questions that were sent by e-mail to participants in two conservation projects in Bangalore, India, to ask for the primary motivations to participate, but also used focus groups and asked the staff about the motivations of their volunteers. The primary motivations reported in their study were 'to protect wildlife', 'to give something back to society' and 'to learn something about wildlife', but also 'to spend time in nature'. In the online project Galaxy Zoo, Raddick et al. (2010) used three experts to identify 12 categories of motivation based on interviews and open questions. In a follow-up study, these categories were then rated on a Likert-scale and extended with new categories from open questions based on a survey with 11'000 participants; 'contributing to science' was the primary motivation for almost 40% of the Galaxy Zoo (Raddick et al., 2013).

In the growing body of literature on the motivation of citizen scientists, there are different ways to classify and summarize motivations based on quantitative surveys or interviews (i.e., different categorization schemes). The theoretical background on motivation in the field of citizen science has often been drawn from psychology and/or the literature on volunteering. Actually, before being called 'citizen science', many of these projects were labelled 'volunteering-projects' (Roy et al., 2012) and citizen scientists can often be considered volunteers in a scientific project. For example, West and Pateman (2016) brought together several theories from the volunteering literature (Clary and Snyder, 1999; Finkelstien, 2009; Locke et al., 2003; Penner, 2002) to describe the factors that influence participation in citizen science. They used, for instance, intrinsic and extrinsic motivation (Finkelstien, 2009) as two overarching categories, which contained the six categories of the 'functional approach to volunteering' by Clary and Snyder (1999). Frensley et al. (2017), used the psychology-grounded self-determination theory, which is based on the three psychological needs of competence, relatedness, and autonomy (Ryan and Deci, 2000a), to categorize and explain participants' motivations in the Virgina Master Naturalist programme (http://www.virginiamasternaturalist.org). Alternatively, Beza et al. (2017) manually extracted seven motivational factors from the citizen science literature and grouped them according to the framework of motivations that lead to community involvement of Batson et al. (2002), which consists of five motives: altruism, collectivism, principlism, intrinsic egoism, and extrinsic egoism (see also section 1.2.1). Finally, Levontin et al. (2018) conducted a more comprehensive literature study on the motivations in a multitude of citizen science projects and reformulated the answers from these studies into 58 statements. These statements were then grouped into 16 categories of personal values[1], which encompass the entire spectrum of human

[1] Version of the questionnaire published in March 2018 (see supplemental materials).

motivation defined by Schwartz et al. (2012) (see also section 2.3.2).

Most studies focus on a single project and use only one scheme to classify the results. The different approaches and surveys to assess the motivations of participants, the different schemes to classify the motivations with different levels of detail, and the substantial differences in the projects make it difficult to compare the results of the different studies on motivations to participate in citizen science projects. Here, we aim to expand the knowledge on the motivation of citizen scientists by comparing two smartphone-based, environment and outdoor focused projects in Europe: CrowdWater (www.crowdwater.ch, Kampf et al., 2018; Seibert et al., 2019b, 2019a) and Naturkalender (www.naturkalender.at). Both projects use smartphone apps based on the SPOTTERON platform (www.spotteron.net) and are available for Android and iOS. The aim of the CrowdWater project is to collect hydrological data, such as water levels, soil moisture and the status of temporary streams. The Naturkalender project (English: Nature's Calendar) focuses on documenting the phenology of indicator species and changes related to climate change. The two projects have, so far, mainly recruited participants from western European countries (most of the participants come from Switzerland and Austria). The comparison of the motivations to participate in the two projects enables a more explicit focus on how the project topic, thematic content and outreach activities affect the motivations of the participants because the projects are similar in terms of the visual design of the app, the way data are transmitted, and the cultural background of the participants. The goals of this study were (i) to identify the motivations of citizens to join the CrowdWater or Naturkalender projects and to see whether these motivations were fulfilled by their participation, (ii) to determine if the main motivations to participate differ for the different demographic groups or between participants who contribute frequently and those who contribute occasionally, and (iii) to contribute to the understanding of motivations to participate in citizen science projects in general. We classified the statements in the questionnaire according to the scheme of Batson et al. (2002), which was adapted by Beza et al. (2017) and is hereafter referred to as “Batson-scheme”, to obtain an overview of the broad categories of motivation. Additionally, we used the scheme of Schwartz et al. (2012), which was adapted for citizen science projects and recently published in a questionnaire by Levontin et al. (2018), hereafter referred to as "Schwartz-scheme", to gain more detailed insights for the entire spectrum of motivations.

## 1.1 The CrowdWater and Naturkalender projects

We investigated the motivations of participants in two smartphone-based citizen science projects, CrowdWater and Naturkalender, which both focus on the environment. The smartphone applications (hereafter referred to as 'apps') for both projects were developed in close collaboration with SPOTTERON, an Austrian company specialised in the development and maintenance of apps for citizen science projects. Each app user can start observations at a new spot and contribute observations to existing spots (i.e., those started by other users) to obtain a time series of observations. The apps include social media functions that enable interaction between participants, such as following other participants, commenting, and liking contributions (Figure 1).

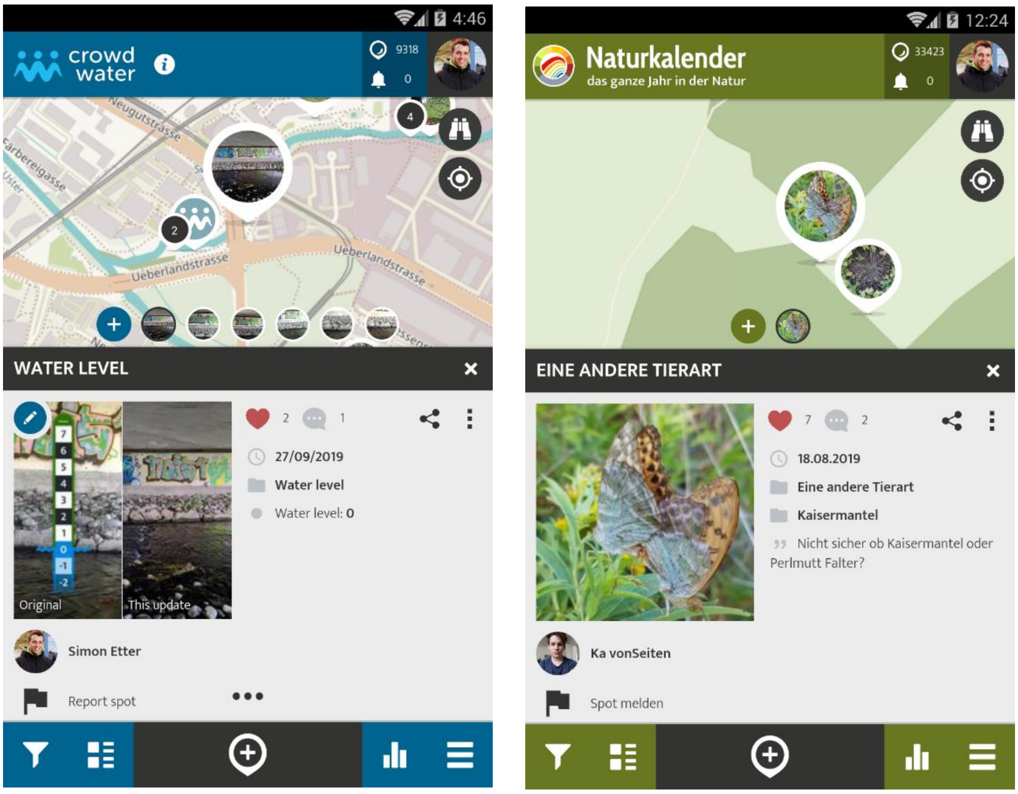



*Figure 1 Screenshots of the CrowdWater (left) and the Naturkalender app (right), with on the top row of the second panel the social media features (from left to right the like button and counter, the speech bubble that allows users to comment on the observation (with the counter next to it), and the sharing button to share contributions on Facebook, Twitter and Google+. More information on the app design can be found in* Seibert et al. (2019a, 2019b) *and spotteron.net.*

### 1.1.1 CrowdWater

The CrowdWater project (www.crowdwater.ch) started in 2016; the app was launched in early 2017. The goal of the project is to develop a tool to collect hydrological information for models that can be used for flood warning and other water management applications. Citizen scientists are asked to contribute pictures of streams and to estimate water level classes based on a virtual staff gauge (Seibert et al., 2019b, 2019a), or to estimate soil moisture based on qualitative classes (Rinderer et al., 2012), or to determine the state of temporary streams (Kampf et al., 2018). Citizen scientists are encouraged to make repeated observations at a location to obtain time series for that location. Observations can – and have been made – around the globe. However, most of the advertisement and outreach activities so far focused on German speaking citizens; hence most observations have been made in Switzerland and Austria.

Social interaction in the CrowdWater app occurs mainly between the project team and citizen scientists via the comments function or by personal communication via e-mail. Only in rare cases do citizen scientists comment on each other's observations. The CrowdWater project has so far mainly been advertised via social media, our private and work-related networks (e.g., presentations at conferences, schools and science fairs, articles in university newsletters and magazines, etc.). Since the value of the data is still subject to research, communication regarding the potential use of the data (e.g. for flood warning systems) has been done very carefully. At the end of October 2018, when the questionnaire was closed, there were 265 users who contributed at least one observation via the CrowdWater app; there were on average 132 contributions per month between February 2017 and October 2018.

### 1.1.2 Naturkalender

Naturkalender (in English: Nature's Calendar) (www.naturkalender.at) is a citizen science project that aims to document the phenology of several indicator plant species throughout the year, to record the behaviour of wild animals, and to document winter phenomena, e.g. the presence or absence of snow cover. By observing the start of, for instance, leaf development or the return of birds from their winter habitats, the project aims to assess the influence of climate change on flora and fauna. Citizen scientists can report the state of plant growth and behaviour and presence of birds, butterflies and bees on a map that covers the entire globe. However, most contributions have been made in Austria. The data collected using the app are included in the Pan European Phenology Project PEP725-database (www.pep725.eu). Naturkalender started in 2014 and was first called "NaturVerrückt". The project consists of multiple apps focusing on different parts of Austria (www.naturkalender.at/regionalprojekte). We sent the questionnaire to the users of all Naturkalender apps and for brevity refer to them as the Naturkalender App.

The Naturkalender app contains a lot of information about plant species and birds, butterflies and bees. Compared to the CrowdWater community there is more communication between participants in Naturkalender. Many observations are commented on by different users, and users help each other with the identification of species. At the time that the questionnaire closed, there were 642 users who provided at least one contribution; there were on average 422 contributions per month between April 2015 and October 2018.

## 1.2 Frameworks

The two frameworks used in this study differ in their origins and foci: The Batson-scheme was designed to describe motivations for community-involvement, while the Schwartz scheme was originally designed as a model of human values. Schwartz (1992) defined values as overarching goals that vary in importance and that serve as guiding principles in the life of a person. These, therefore, have a strong influence on the motivations of individuals. The Batson-scheme has already been used for citizen science projects (Beza et al., 2017), whereas the Schwartz-scheme had not been used for motivations in citizen science when this study was conducted. The combined use of two frameworks allows interpretation of more differentiated results from the same set of questions. For instance, the Batons-scheme explicitly distinguishes between egoistic and non-egoistic motivations. As can be seen later, the categories of the Batson-scheme represent the individual statements in a category more reliably. In contrast, the Schwarz-categories provide more detailed insights, but the results are overall less reliable.

### 1.2.1 Batson-Framework

Batson et al. (2002) offer a framework to classify motivations for community engagement based on four categories: *egoism*, *altruism, principlism and collectivism. Egoism* describes the motivation of a person who seeks primarily his/her own benefit in doing something. The actions taken might still serve the community or the greater good, e.g. volunteering in a citizen science project in order to be able to include that in one's résumé. *Altruism* is defined as the motivation to fulfil someone's needs and is mostly motivated by the feeling of empathy towards the other person. An example is to volunteer in a citizen science project to help researchers with their work. *Collectivism* is the motivation to increase the welfare of a group, e.g. by measuring and reporting lead pollution in tap water of the local community, as in Pieper et al. (2018). *Principlism* is defined as the motivation to uphold some moral principle(s), like justice or the conservation of wildlife (Batson et al., 2002).

The framework of Batson et al. (2002) has been applied to citizen science by Beza et al. (2017). They combined it with the framework of Ryan and Deci (2000a) to distinguish intrinsic egoism *(egoism, intrinsic)* focused on a person's satisfaction (e.g., fun or interest in sharing information) and extrinsically motivated egoism (*egoism, extrinsic*) that aims to achieve a desirable and separate outcome (e.g. expecting something in return). We chose this framework because it provides a good overview of the motivations of the participants with relatively simple and easily interpretable categories. The attribution of the statements used in the survey to these five categories can be found in Table S1.

### 1.2.2 Schwartz-Framework

To use the findings of questionnaires on motivation to improve the design of citizen science projects, it is beneficial to use a framework that encompasses the entire spectrum of motivations and enables a more detailed assessment of the motivations. Schwartz et al. (2012) developed a framework of 19 basic values based on the values described by Schwartz (1992). These values express the guiding prin-ciples in a person's life and form the base of the person's decisions. The values are distributed in a circular continuum (Figure 2) with the four dimensions: self-enhancement (improving oneself) and its counterpart self-transcendence (investing in other people/things), conservation (preserving the status quo) and its counterpart openness to change (Schwartz, 1992). Levontin et al. (2018) adapted this framework slightly to make it suitable for citizen science projects. This resulted in 16 values (as of March 2018), which are, hereafter, referred to as categories. The attribution of the statements used in the survey to these 16 categories can be found in

Table S2.

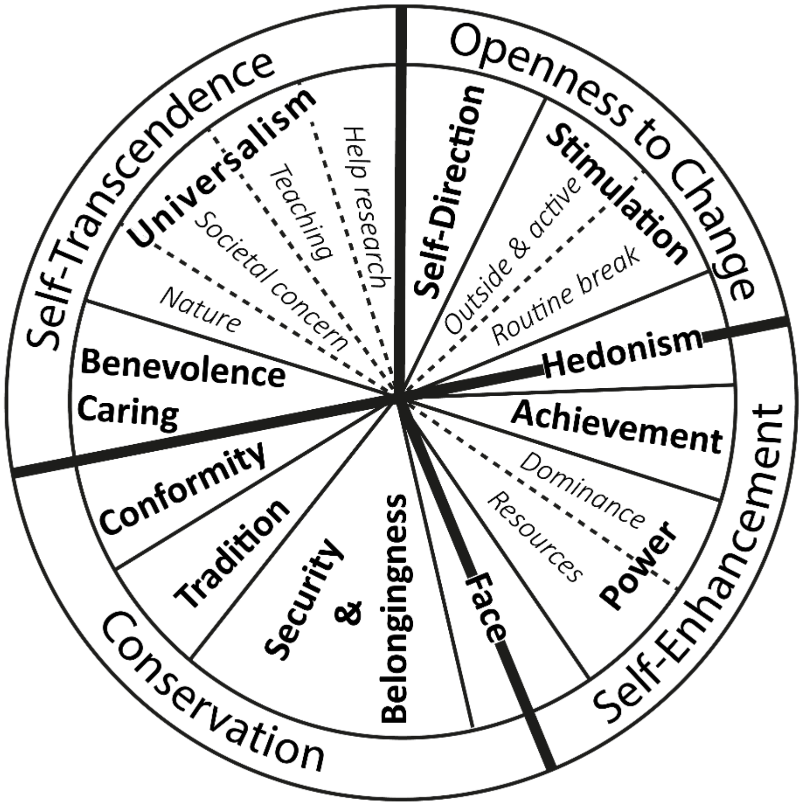



*Figure 2 The circular continuum of personal values from Schwartz et al. (2012) adapted using the category names of the questionnaire designed by Levontin et al. (2018) for citizen science projects. All categories in bold font in the inner circle and their subcategories (in italic) reflect one or multiple statements in the questionnaire used in this study. The description of the categories can be found in Table S2.*

# 2 Methods

## 2.1 Questionnaire

In the first part of the questionnaire, the engagement part, we aimed to identify the motivations of citizen scientists that led to their engagement in either the CrowdWater or Naturkalender projects. Based on the definition of motivations of Phillips et al. (2018), we interpreted the motivations to become engaged in a project as goals that can potentially be fulfilled by participation. In the second part of the questionnaire, the fulfilment part, we aimed to see which of these initial motivational goals were fulfilled by participation in the projects. Some of the statements in the fulfilment part are related to the construct of (self-)efficacy, which refers to a person's belief of being capable to learn specific things or to perform particular actions (Bandura, 1997; e.g. “By doing this activity I can help others”). However, not all the statements overlap with the above definition (e.g. “This activity is fun for me” or “This activity increased my social status”). Therefore, we use the term fulfilment throughout the text, even when it refers to efficacy. We selected 29 of the 58 statements of the questionnaire that was developed during a citizen science COST action workshop[2] in Latvia in March 2018, and published by Levontin et al. (2018), e.g. “I participate in the project because I want to do something meaningful (see Supplementary Material 1 for the questionnaire). We asked the participants in how far they agreed with these statements based on a five-point Likert-scale with the options “don’t agree at all”, “rather don’t agree”, “undecided”, “rather agree”, “fully agree”. Most statements were rephrased to make them more suitable for the fulfilment part. It was, however, not possible to rephrase all of them in a meaningful way. This was for example the case for the statements of the categories *conformity* (trying to act in a way that does not harm or upset anyone and fulfils social expectations or norms (Schwartz et al., 2012).) and *power, resources* (maintaining or achieving social status and prestige by controlling or acquiring resources; Schwartz et al., 2012). Furthermore, to avoid confusion we decided to leave out “I enjoy this activity” in the engagement part because we assumed that participants of CrowdWater were very unlikely to have participated in hydrological data collection before initial participation in the project and thus cannot reliably state that they already enjoyed this activity before participating in the project.

An invitation to fill out the online questionnaire on surveymonkey.com (in English and German) was sent to all participants of the two projects on August 8th, 2018 with push messages in the apps and by e-mail to the 400 people who had registered for the CrowdWater newsletter at that time. Only the participants of the CrowdWater project were reminded by a second push message on August 22nd, 2018.

[2] https://cs-eu.net/news/workshop-report-wg-4-motivation-participants-citizen-science-projects

## 2.2 Analyses

We classified the statements in the questionnaire using the categories of the Batson-framework (Table S1) and those of the Schwartz-framework (

Table S2). For each statement, we determined the percentage of respondents who agreed (i.e., those who chose “rather agree” or “fully agree”) with the statement. We also determined the average percentage of respondents who agreed with the different statements in each category of either the Batson or Schwartz framework. For categories with more than two statements, we used Cronbach’s alpha (Cronbach, 1951) to assess the consistency of the agreement to the different statements in a category (i.e. a reliability analysis). For the categories with only two statements, we used the Spearman-Brown coefficient (Eisinga et al., 2013). To avoid a lengthy questionnaire and due to the inability for some statements to be used in the engagement or the fulfilment part, there were several categories in the Schwartz-categorisation that had only one statement per category. For the categories with only one statement, the calculation of a reliability (or consistency) score is not possible, nor necessary.

To determine the statistical significance of differences in the agreement with statements, the answers to the statements in the questionnaire were converted into numbers from 1 to 5: 1 for “don’t agree at all”, 2 for “slightly disagree”, 3 for “undecided”, 4 for “slightly agree” and 5 for “fully agree”. We used the paired Wilcoxon signed rank test to test the significance of the differences in the median response to the statements regarding the motivations for initial engagement and the fulfilment of these motivations by participating. We used the Mann-Whitney U-test to test the significance of the differences in the median response for different subgroups of respondents (e.g. CrowdWater vs. Naturkalender participants, super-users vs. occasional participant, the different age groups, etc.). We used a significance value of 0.05 for all analyses.

# 3 Results

## 3.1 Number of Responses and Demographics

We received 101 responses, but only 90 could be used in this study. We excluded answers from people who never contributed to the project (some of the people who subscribed to the CrowdWater newsletter had never used the app), incomplete questionnaires, as well as answers from people who work for one of the projects or SPOTTERON. Of the 90 questionnaires with complete responses, 54 were submitted by CrowdWater participants and 36 by Naturkalender participants.

Based on the 400 emails and 265 active participants in CrowdWater and 642 in Naturkalender, we estimate a response rate of about 8%. This number is, however, highly speculative as people might have uninstalled the app before we sent the push message. Furthermore, people who had installed the app but had never contributed might have received the invitation but were not counted by us. Based on Israel (1992), the number of responses in each project, and the comparison with the assumed number of active participants (265 in CrowdWater and 642 in Naturkalender), this survey is a convenience sample.

We have no data to determine the representativeness of the respondents for the participants in the projects but assume that they either represent the participants or include more frequent users. Most of the respondents (n=25) were in the 30-39 age group. There was a gender balance for the respondents (54% female vs 46% male) but it is unknown to what extent this reflects the participants in the projects because neither of the projects records the gender of the participants. For many environment-related volunteering or citizen science projects (Geoghegan et al., 2016; Raddick et al., 2013; Wright et al., 2015) and outdoor projects (Alender, 2016; Land-Zandstra et al., 2016a) there is a slight overrepresentation of male participants (often between 50 and 60 %). On the other hand Land-Zandstra et al. (2016b) report that more females participated in the Dutch flu-tracker project (55%) and Pandya and Dibner (2018) report that, by the end of 2017, 65% of the user profiles on the citizen science platform SciStarter (scistarter.org) were created by females. Based on our experience, the distribution of female and male participants is fairly balanced. Furthermore a study in Switzerland indicated that gender is not a significant indicator of interest in citizen science (Füchslin et al., 2019).

*Table 1 Number of respondents by gender and age group and number of super-users and occasional participants for the two projects. Super-users are users who said that they contribute at least one observation per week.*

| | Gender | Project | Frequency of contribution |
|---|---|---|---|

| **Age group** | Female | Male | CrowdWater | Naturkalender | Super-users | Occasional participants |
|---|---|---|---|---|---|---|
| **<18** | 3 | 1 | 3 | 1 | 1 | 3 |
| **21-29** | 12 | 4 | 13 | 3 | 1 | 15 |
| **30-39** | 13 | 12 | 18 | 7 | 7 | 18 |
| **40-49** | 6 | 7 | 9 | 4 | 4 | 9 |
| **50-59** | 10 | 9 | 7 | 12 | 11 | 8 |
| **60+** | 8 | 4 | 4 | 8 | 6 | 6 |
| **not stated** | 1 | | 0 | 1 | 1 | 0 |
| **Total** | 52 | 37 | 54 | 36 | 31 | 59 |

We classified respondents who stated that they contribute to the projects at least weekly as super-users (n=31, Table 1) and all other users as occasional participants. There were 14 super-users for CrowdWater and 17 for Naturkalender (Table 2). Eleven out of the 31 super-users (35%) were between 50-59 years old.

*Table 2 Number (and percentage) of respondents that are super-users and occasional participants for the CrowdWater and Naturkalender projects.*

| | **Super-users** | **Occasional participants** |
|---|---|---|
| **CrowdWater** | 14 (26%) | 40 (74%) |
| **Naturkalender** | 17 (47%) | 19 (53%) |
| **Total** | 31 (34%) | 59 (66%) |

## 3.2 Consistency of the results for the different categories

The consistency of the agreement to the different statements in a category (i.e., the reliability of the category) can be considered "good" or "acceptable" for all Batson categories (Cronbach's alpha > 0.7; George and Mallery, 2003) with more than two statements (Figure S1). The category altruism, which only included two statements, had a Spearman-Brown score of 0.64 for the engagement part and 0.53

for the fulfilment part, which indicates a "questionable" and a "poor" consistency but is still somewhat acceptable according to George and Mallery (2003).

For nine out of the 16 Schwartz-categories the Cronbach's alpha or Spearman-Brown-score was higher than 0.5 for the engagement part and it was higher than 0.5 for seven out of 14 categories in the fulfilment part. For the Schwartz categories with three statements (the maximum number of statements per category), the Cronbach's alpha was larger than 0.5, except for the category *power, dominance* (maintaining social status and prestige by controlling and dominating other people; 0.44) in the engagement part and *achievement* (achieving goals according to social standards and thereby demonstrating competence; 0.48) in the fulfilment part (the explanations in parentheses are based on Schwartz et al., 2012). According to the Spearman-Brown test, the reliability for the two-statements in the categories, *benevolence, caring* (improving or preserving the wellbeing of people that are relevant in one's everyday life; 0.48), *face* (security and power by avoiding humiliation and maintaining a good reputation; 0.34), and *stimulation and routine break* (doing exciting and new things that might also challenge oneself; -0.47) in the engagement part was poor. For the fulfilment part, the reliability for the categories *benevolence, caring* (0.23), *self-direction* (independent exploring, learning and being creative; 0.46), and *universalism, nature* (upholding the value of nature and protecting it; 0.41) was poor. Even though, the reliability analysis indicates that not all categories can be considered a reliable representation for all the statements in the category, we still describe the results of the questionnaire mainly per category, rather than per statement, to highlight the main results. For the categories with low reliability, we also report the agreement for the individual statements.

## 3.3 Motivations for Initial Engagement in CrowdWater and Naturkalender

The median agreement to statements was significantly higher for the Naturkalender respondents for both initial engagement (median 4 – "rather agree" for Naturkalender vs. 3 – "undecided" in CrowdWater) than for the CrowdWater respondents. A*ltruism* was the main motivational factor according to the Batson-scheme to join CrowdWater (i.e., it was the factor with the highest average agreement; 82%), whereas for Naturkalender it was *principlism* (89%; Figure 3; see Figures Figure S2 and Figure S3 for the agreement to the individual statements). The order of the categories with the highest average agreement didn't differ between the two projects for any of the other categories. However, Naturkalender respondents agreed significantly more with the Batson categories *egoism-intrinsic, collectivism*, and *principlism* than the CrowdWater respondents (all p-values<0.01).

The four Schwartz-categories with the highest average agreement were the same for CrowdWater and Naturkalender, but the average agreement was again higher for the Naturkalender respondents than the CrowdWater respondents (Figure 3; see Figures Figure S4 and Figure S5 for the agreement to the individual statements). These top categories were (with explanation according to Schwartz et al., 2012): *universalism, help with research* (upholding the value of science and support it; 90% agreement for CrowdWater vs 94 % for Naturkalender), followed by *universalism, nature* (83 vs 94 %), *self-direction* (81 vs 88 %) and *universalism, societal concern* (appreciating the value of society, protect and improve it 78 vs 85 %). For the categories for which the average agreement was lower, the order of agreement differed somewhat between the CrowdWater and Naturkalender respondents (Figure 3). The CrowdWater respondents agreed significantly more to statements related to *conformity* (47 vs. 17 %, p<0.01) and *stimulation and routine break* (42 vs. 24%; p=0.02) than Naturkalender respondents. CrowdWater respondents agreed significantly less with statements related to *universalism*-teaching (upholding the value of teaching and sharing experiences;; 53 vs. 70 %; p=0.02), *security and belongingness* (safety by feeling connected to a community; 34 vs. 50 %; p<0.01) and *stimulation-being outside and* active (49 vs. 80 %; p<0.01).

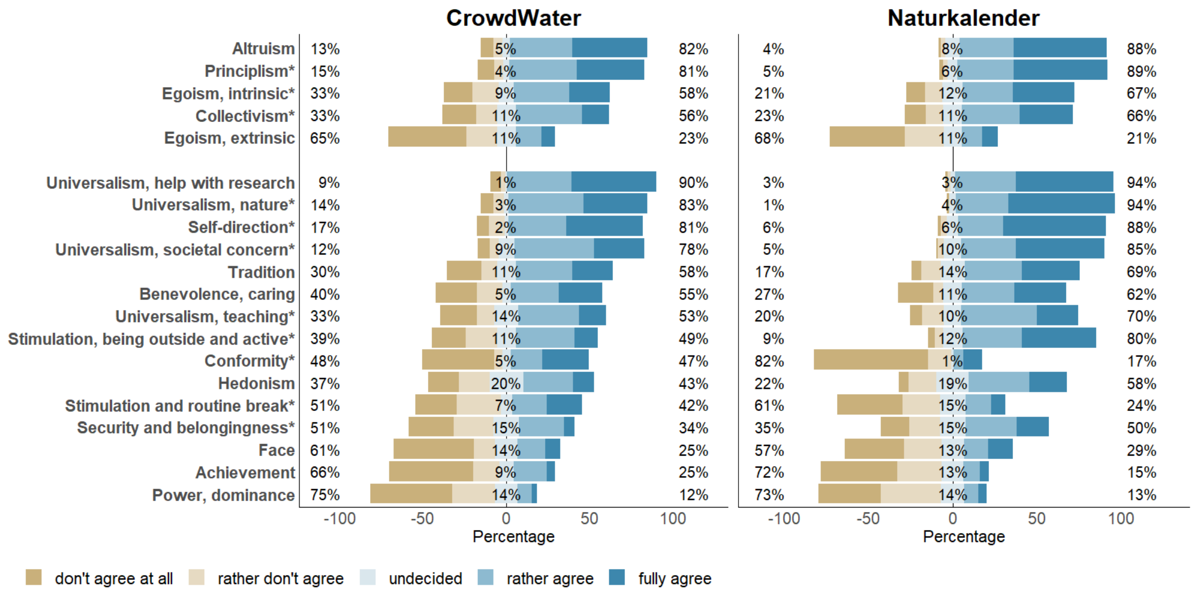



*Figure 3 Percentage of respondents who chose one of the five levels of agreement to statements regarding initial engagement that belong to the motivational categories according to Batson et al. (2002) (top five rows) and Schwartz et al. (2012) for CrowdWater (left) and Naturkalender (right) . For the categories marked with an asterisk (*), the median response for the CrowdWater and Naturkalender respondents were significantly different. The values next to the categories indicate the percentage of respondents who don't agree (left; don't agree at all and rather don't agree), are undecided (middle) and agree (right; rather agree and fully agree). The categories are sorted by decreasing percentage of agreement for the respondents of the CrowdWater project. Figures Figure S2-Figure S5in the supplemental material show the percentage of agreement for the individual statements in each category. .*

## 3.4 Fulfilment of Motivations in CrowdWater and Naturkalender

The top motivational factors that were fulfilled by participating in the projects according to the Batson-scheme were *altruism*, *principlism* and *egoism-intrinsic* for both the CrowdWater and Naturkalender respondents (Figure 4; see Figures Figure S6 and Figure S7). Even though for *principlism* the average agreement was 81 % for both projects, the median response for the Naturkalender respondents was significantly higher due to the larger percentage of Naturkalender respondents who fully agreed with these statements (37 % for CrowdWater vs. 23 % for Naturkalender, p=0.02). Naturkalender respondents also agreed significantly more to motivation factors in the *egoism-intrinsic* category, again due to a higher percentage of respondents who fully agreed with these statements (21 % for CrowdWater vs. 33 % for Naturkalender respondents, p<0.01). These differences can be attributed to the very high agreement (92 % or more) of the Naturkalender respondents to the statements “By contributing to this project I can share my knowledge and experiences”, “I enjoy this activity”, “This activity taught me new skills or new knowledge" and "This activity is fun for me".

Compared to the motivations for initial engagement of the Naturkalender respondents, a significant decrease was observed in the median responses to the statements in the categories *altruism* (p<0.01), *collectivism* (p=0.04) and *principlism* (p<0.01), and a significant increase for the category *egoism-intrinsic* (p=0.02). The CrowdWater respondents agreed significantly less with the categories *collectivism* (p<0.01) and *egoism-extrinsic* (p=0.02) compared to the agreement for initial engagement.

The average agreement with the statements in the categories *universalism-help with research* and *universalism-nature* in the Schwartz's scheme remained high after initial participation for the respondents of both projects (Figure 4; see Figures Figure S8 and Figure S7). The median agreement for the statements in the categories *hedonism* (experience pleasure and enjoyment physically or mentally; Schwartz et al., 2012) and *achievement* increased significantly after participation for both projects (all p-values<0.01). For Naturkalender respondents, *hedonism* was the category with the highest agreement (it was ranked 9$^{th}$ in the engagement part; Figure 3). Significantly fewer CrowdWater respondents agreed to statements related to *hedonism* (p<0.01), so that it was the 4$^{th}$ ranked category based on the percentage of agreement (Figure 4). The category with the second highest agreement for Naturkalender respondents was *self-direction* because 97% of the respondents agreed with the statement "This activity taught me new skills or knowledge". For the CrowdWater respondents, the agreement to this category was much lower (67 % agreement, ranked 6$^{th}$) and also much lower than for the initial engagement (81 % agreement, ranked 3rd). The median response for *self-direction* was 4 (rather agree) for both projects but the percentage of respondents who fully agreed with the statement "This activity taught me new skills or knowledge" was much lower for CrowdWater respondents than for the Naturkalender respondents (18 vs. 34 %), which made the difference statistically significant (p<0.01).

The average agreement to the statements in the category *tradition* (upholding traditional principles, values, and customs of a culture or religion; Schwartz et al., 2012) increased compared to the initial motivation for engagement for the CrowdWater respondents (becoming the category with the third highest agreement, although the difference in the median response for initial engagement and fulfilment was not statistically significant; p=0.11). For Naturkalender respondents, the average agreement for this category barely changed compared to the agreement for initial motivations. The agreement in the following categories decreased significantly compared to the initial engagement: *self-direction* (CrowdWater only, p<0.01), *universalism-societal concern* (both projects, both p-values<0.01), *stimulation, being outside and active (Naturkalender only, p<0.01), security and belongingness* (both projects, both p-values<0.01), *and face* (both projects, both p-values<0.01). The three categories for which the average agreement was the lowest were *face, security and belongingness* and *power, dominance* for both projects (Figure 4).

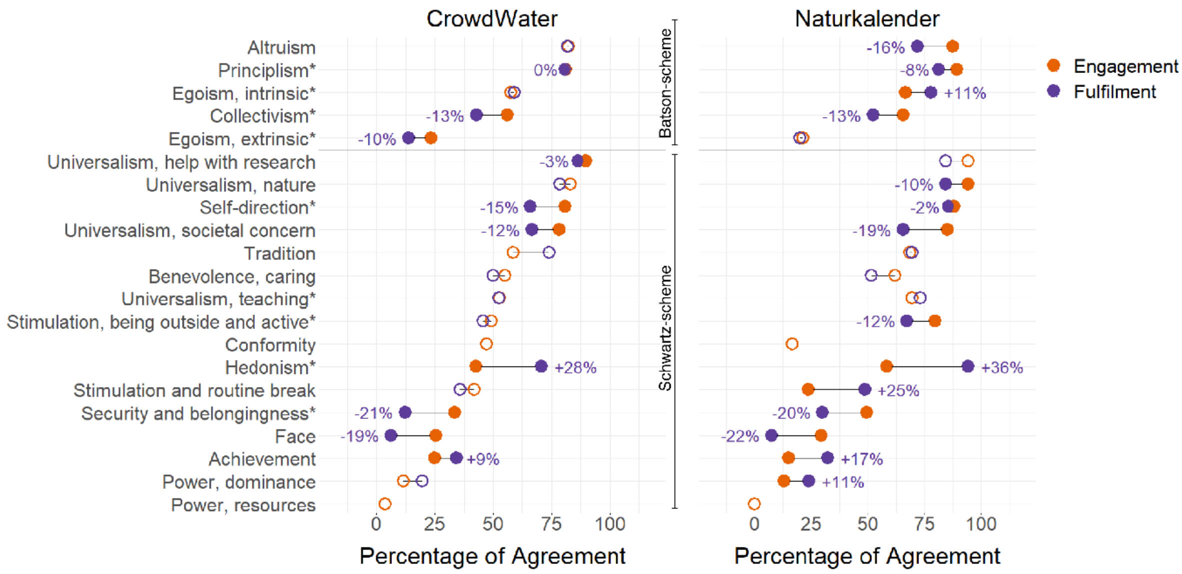



*Figure 4 The average percentage of respondents that agreed to the statements that belong to the different categories for the motivations for initial engagement (orange) and fulfilment (purple) for CrowdWater (left) and Naturkalender (right). Empty circles indicate insignificant (p>0.05) changes in the median response for initial engagement and fulfilment; filled symbols indicate significant changes. Asterisks indicate categories for which the median response for fulfilment for the CrowdWater and Naturkalender respondents was significantly different (see Figure 3 for the statically significant differences in the agreement for initial engagement). The categories are sorted by decreasing percentage of agreement for the CrowdWater respondents in the engagement part. Figures S5-S9 in the supplemental material show the percentage of agreement for the individual statements per category.*

## 3.5 Super-Users vs. Occasional Participants

For the initial engagement, the super-users agreed significantly more to statements related to *egoism-*

*intrinsic* and *principlism* in the Batson-scheme than the occasional participant (68 vs. 58% and 86 vs. 83%, respectively (p<0.01 for both); Figure S10). For the categories in the Schwarz-scheme, the super-users agreed significantly more than the occasional users to *universalism-help with research* (93 vs 91%; p<0.01) and s*elf-direction* (88 vs. 82%; p=0.04), *stimulation-being outside and active* (72 vs. 55%; p<0.01) and *security and belongingness* (50 vs 34%; p<0.01).

There were also some differences among super-users and occasional participants within the projects: For the CrowdWater project, the occasional participants were significantly more motivated to join the project by *conformity* than the super-users (56 vs. 22%, p=0.03). The difference between the occasional participants in Naturkalender and the occasional users in Naturkalender was also significant (56 vs. 13%, p<0.01). For the Naturkalender project, there was no significant difference in the median response for the statements related to *conformity* for the super-users and occasional participants (21 vs. 13%, p=0.80).

The agreement to statements related to the fulfilment of the motivations was generally higher for super-users than for the occasional participants, but the ranking of the categories to which the respondents agreed most was very similar. The differences in the median response of the super-users and occasional participants were statistically significant for the same categories as for the initial engagement (i.e., *egoism-intrinsic* and *principlism* (Batson-scheme), *universalism-help with research*, s*elf-direction*, *stimulation-being outside and active*, and *security and belongingness* (Schwarz-scheme)), but for the fulfilment super-users also agreed significantly more to statements related to *power, dominance* (27 vs. 18%, p<0.01) and *achievement* (40 vs. 30%, p=0.01).

## 3.6 Age

In the fulfilment part, the respondents younger than 50 agreed most to statements related to *altruism* (83-88%) and second most to statements related to *principlism* (79-88%), whereas the 50-59-year old respondents agreed most with statements in the *egoism, intrinsic* (78%) and *principlism* (78%) categories. The respondents above 60 years agreed most to statements in the *principlism* category

(77%).

For the fulfilment part, the age group 50-59 was the only group that agreed most to the category *hedonism* (89%) in the Schwartz-scheme. In contrast, respondents in the other age groups agreed most to *universalism, help with research* or *universalism, nature*. Furthermore, the respondents in the 50-59 age group agreed significantly more to statements related to *stimulation, being outside and active* (71%, p=0.01; doing exciting, new and challenging things in the outdoors and being physically active; Schwartz et al., 2012) than the respondents in all other age groups combined (49-71%). On average, the respondents in the 50-59 age group also agreed more than other age groups to statements in *stimulation, being outside and active* for the initial engagement (84 %, vs 68 % or less for the other age groups).

# 4 Discussion

## 4.1 Limitations of the Study

The reliability of the grouping of the statements into the categories of Batson et al. (2002) was satisfactory but the reliability was poor for several categories in the Schwartz-scheme. We removed some statements due to a necessary trade-off between a lengthy questionnaire and more statements per category in order to be able to include questions regarding the engagement and fulfilment. This probably impacted the reliability of the categories, and it remains necessary to test if the reliability of the categories is higher for different projects, other geographic settings or with a different selection of statements.

The convenience sample is also a limiting factor of this study. The respondents might not fully represent all participants in CrowdWater and Naturkalender. More engaged participants, for instance, may have been more likely to fill in the questionnaire. Furthermore, biases like the social desirability bias (Furnham, 1986), where people give answers that are not necessarily true but that they think are socially more desirable, cannot be excluded entirely either. However, impersonal and anonymous distribution of questionnaires (as in this study) reduces this social desirability bias (Nederhof, 1985).

We, therefore, assume that the results from the questionnaire provide useful information on the main motivations to initially participate and to continue participating in the CrowdWater and Naturkalender projects.

## 4.2 Motivations for Initial Engagement

We evaluated to which extent the motivations of the participants to join the CrowdWater and Naturkalender projects agreed with the motivational factors mentioned in the peer-reviewed literature (Levontin et al., 2018). The similar order of the percentage of agreement for the motivations to engage in CrowdWater and Naturkalender suggests that people had similar expectations prior to participation. The participants of both projects expected to contribute to science, to protect nature, to learn something new, but also to satisfy their interest in the topic, and to address social concerns. This is in line with Alender (2016), who found similar motivations for participants of eight water quality monitoring projects in the US. De Vries et al. (2019) reported, based on a literature review across multiple projects in the natural sciences and health, that helping science is an important motivation as well. To help science or help with research was also a main motivation for online projects, such as Foldit (Curtis, 2015) and Galaxy Zoo (Raddick et al., 2013), for aerosol monitoring (Land-Zandstra et al., 2016), and for flu reports using smartphones (Land-Zandstra et al., 2016b).

The high agreement to motivations related to *universalism, nature* for Naturkalender and CrowdWater suggests that protecting the environment is an important issue for the participants. For environment-related citizen science projects the topics or issues addressed by the project, or protecting the environment in general, are often important motivational factors (Alender, 2016; Hobbs and White, 2012; Johnson et al., 2014; Ryan et al., 2001). For example, Hobbs and White (2012) found that for several British wildlife-conservation projects, the interest in wildlife and the contribution to wildlife conservation were the two main motivations to join the projects.

Land-Zandstra et al. (2016b) found that learning, fun or socializing were weak motivators to become involved in the flu-tracker project. We could confirm this for fun (*hedonism*) and socializing (*security* *and belongingness*) for Naturkalender and CrowdWater, but not for learning (*self-direction*). A reason for this discrepancy could be that there were probably few learning opportunities for participants of the flu-tracker project because they only report flu symptoms (Land-Zandstra et al., 2016b). In Naturkalender, participants can learn about phenology and in CrowdWater to a lesser degree about fluctuations in water levels in the observed streams. In this respect, the difference between CrowdWater, where other than deliberately observing hydrological changes there is no learning involved, and Naturkalender, where participants learn to identify particular species, is interesting. The difference in the agreement that learning (*self-direction*) was a motivation to join the project for the two projects was small (81% for CrowdWater vs. 88% for Naturkalender), suggesting that participants for both projects wanted to learn something. The difference in the agreement that the projects fulfilled the learning motivation was indeed much larger (66% for CrowdWater and 86 % for Naturkalender (86 %).

Socialising aspects, i.e., meeting new people were not a strong motivator, which might be due to their app-based character of the CrowdWater and Naturkalender projects. Participants typically have no opportunity to meet each other. We agree with the assumption of Land-Zandstra et al. (2016b) that the type of project makes a difference in this case, namely whether the project offers opportunities to learn and also if they are based on (real-world) social interactions.

## 4.3 Differences in How Far the CrowdWater and Naturkalender Projects Fulfilled the Expectations

### 4.3.1 Learning, Teaching and Social Interactions

The significantly higher agreement to *stimulation, being outside and active* for the Naturkalender respondents than the CrowdWater respondents suggests that Naturkalender participants value being outdoors, in nature and doing a physical activity more than CrowdWater participants. Based on the multitude of existing animal or plant phenology projects that involve volunteers (Beaubien and Hamann, 2011; Fuccillo et al., 2015), we assume that activities in the Naturkalender project are more aligned to hobbies than the observation of hydrological variables in the CrowdWater project. This assumption is also supported by the fact that Naturkalender respondents agreed significantly less to *stimulation and routine break* as a motivation for engagement and especially the higher agreement to the statement “I was doing this activity anyways” than the CrowdWater respondents. This indicates that some respondents of the Naturkalender project were already participating in similar activities as part of a hobby and, therefore, did not join the project to do something completely new but instead were able to share their knowledge. The higher agreement to *universalism, teaching* as a motivator for the initial engagement of Naturkalender respondents indeed indicates that participants in Naturkalender value sharing their knowledge more or had more knowledge to share than the CrowdWater participants. Teaching opportunities in the Naturkalender project include helping other participants with species identification via comments in the app. The much more extensive use of the social media features in the Naturkalender app than the CrowdWater app, reflects the fulfilment of this motivation.

The opportunities for teaching are directly related to the opportunities for learning, which is likely why learning (*self-direction)* was the category with second highest agreement for fulfilment for the Naturkalender respondents (although, the agreement that the project fulfilled this criterion was significantly less than the agreement that learning was a motivator to initially join the project). This matches the findings of Rotman et al. (2012), who stated that motivations like personal interest and curiosity were the most influential factors for continued participation in environment-related projects, such as Biotracker (http://www.birds.cornell.edu/citscitoolkit/projects/biotracker-nsf-project/), which collects images of tree leaves to develop an automatic species recognition application and is thus topic-wise closely related to Naturkalender. In Naturkalender, participants can acquire new knowledge about plant and animal species from information in the app and the comments of other participants; the CrowdWater app does not provide such information, which is likely why the agreement with statements in the self-direction category was much lower for the CrowdWater respondents. CrowdWater offers information about hydrology on the homepage and links to an online course called “Water in Switzerland”. However, so far it appears that these options are rarely used, possibly due to them being mentioned on the homepage, rather than in the app. Thus, opportunities for learning in CrowdWater are limited compared to Naturkalender, where users profit from the expertise of other participants.

According to Serret et al. (2019), tools and features that help participants to form a network can be the basis for a self-organized community, where participants correct each other and share their experience. Even though the comment boxes provide such opportunities for teaching and learning, the low agreement to *security and belongingness* for both the CrowdWater and Naturkalender respondents indicates that comments in the app do not fulfil the motivation to socialise with like-minded people enough. However, it also has to be noted that this was not an important motivation to join in the first place. The two projects are, in that perspective, more similar to other smartphone-based projects, like e.g. flu-tracker (Land-Zandstra et al., 2016b) or online projects (Nov et al., 2014) that do not lead to real-world interaction.

### 4.3.2 Enjoyment, Fun and Conformity

Respondents of both projects agreed significantly more to the fun part *(hedonism)* being fulfilled than it being a motivation to join the project initially. This indicates that although the participants did not join the projects for the fun factor, they continue to participate because they (also) enjoy it. Reasons for the higher agreement to enjoyment as a motivator for the Naturkalender respondents than the CrowdWater respondents might be the fact that in the Naturkalender app, there are many more options and locations where one can report observations of plants, animals and winter phenomena. In the CrowdWater app, the observations are restricted to streams and rivers and soil moisture measurements can be taken only at unpaved locations. Therefore, we assume that more potential locations to contribute and more data entry options, together with more virtual social interaction, increase fun and enjoyment, and thereby the overall motivation for participation.

For the CrowdWater project, the occasional participants were significantly more motivated by *conformity* than the super-users. The higher agreement to *conformity* for the CrowdWater respondents could be explained by the fact that the CrowdWater app had been available for only 17 months prior to the launch of the survey. The first contributions for Naturkalender were made 41 months prior to the start of the survey. Therefore, at the time of the survey, CrowdWater probably still relied more on the immediate social network of the project administrators. Naturkalender attracted many participants through press releases and outreach events. These participants obviously feel less obliged to participate. From the higher agreement to *conformity* for the occasional participants in CrowdWater than the super-users, we conclude that asking people, particularly friends, colleagues or family members, to participate leads to a light form of social pressure for people who would otherwise not be motivated to participate. This might lead to increased participation in the beginning of the project but if people don't find something rewarding in the project, e.g. a fun component or a learning outcome, they might soon stop contributing, even though they agree that helping science, society or protecting nature are worthwhile. Although motivations to make the world a better place, make scientific knowledge available to the public and to contribute to the future of humanity (*universalism, societal concern)* were important motivators to join both projects, they were probably too ambitious to be fulfilled for some participants.

## 4.4 Super-Users and Their Motivations

The median age of the super-users (50-59 age group) was higher than for the occasional participants (30-39 years; Table 1). Many other projects report that the majority of participants are 30 years or older (Alender, 2016; Beza et al., 2017; Land-Zandstra et al., 2016b; Pandya and Dibner, 2018). Although the small number of respondents per age group does not allow us to draw many conclusions related to the age of the participants, the hint of older people being more intrinsically (*egoism, intrinsic*) motivated is interesting. A high degree of intrinsic motivation of participants is desirable for citizen science projects because it leads to more and better contributions (Deci and Ryan, 2000). This is largely because most citizen science projects cannot offer any compensation for the contributions. Based on the high agreement to *hedonism*, *self-direction* and *stimulation, being outside and* active, we assume that the projects fulfil some of the intrinsic motivations for participants in the 50-59 age group by providing an opportunity to go outdoors and be physically active as part of a regular routine. In a study among 8245 US citizens above the age of 65, Szanton et al. (2015) found that physical activities were chosen as the favourite leisure activity across all income and racial groups. In the Community Collaborative Rain, Hail, and Snow Network (CoCoRaHS), older participants reported rainfall observations more timely, reliably and over longer periods, and some participants even reported to have incorporated their measurements into their daily routines (Sheppard et al., 2017). Venkatesh et al. (2012) report that older people are more likely to stick to established habits, which might also explain their higher contribution.

The significantly higher agreement of super-users to *achievement* and *power, dominance* than occasional participants for the statements related to fulfilment indicates that the super-users feel that their contributions are seen and valued. This might motivate them to contribute more than others (Nov et al., 2014). There might be a self-energising mechanism here: participants who contribute more, will probably also have received more likes, “Thank You”-comments, recognition and feedback by the project administrators, which then encourages them to contribute more (de Vries et al., 2019). This leads to their “dominance” over other participants in terms of the number of contributions (e.g., a high place on the leader-boards). However, the agreement to statements related to these competitive categories was rather low. It remains unclear if this was due to the low level of gamification, which at the time of the survey consisted only of a simple leader board, or if the respondents did not want any competition. Whether increased gamification increases the agreement to these motivations as proposed in Nov et al. (2014), therefore, remains to be investigated.

## 4.5 Recommendations for Successful Citizen Science Projects

All citizen science projects depend on dedicated participants; communication with the right target audiences is key to success (Parrish et al., 2018). Therefore, it is essential to identify target groups by characterising the motivations of potential participants, and particularly the super-users. Although this depends on the project (including the topic and tasks involved), some general conclusions can be made based on the findings from our survey and those reported in the literature.

- It appears that people are more likely to contribute to a project over extended time periods, if they have shared values with the project's goal (e.g. protection of the environment; see also relatedness to a topic in self-determination theory; Ryan and Deci, 2000b). Moreover, the level of interest increases if projects tackle problems that impact the every-day life of participants (Frensley et al., 2017), such as a local issue of the community (e.g. PublicLab; Rey-Mazón et al., 2018). One could argue that everyone, and thus also Naturkalender participants, is affected by climate change and people can observe the effects in their backyard. For CrowdWater, the local relevance of their stream observations was less evident because the data are not linked to any forecasts (yet). Furthermore, people may expect the government to be responsible for flood or drought forecasting and water management. The motivation to participate in CrowdWater might be different in other countries where people are more exposed to flood hazards.
- Participants need to be interested in the topic of the project and the activities involved. They often have an interest in science or technology. For online projects, the motivation to participate in a project is mainly to contribute to science (Curtis, 2015; Raddick et al., 2013). The agreement to the statement “I am interested in the topic of this project” was very high for respondents for both projects, similar to the findings of Hobbs and White (2012) for two wildlife observation projects. For Naturkalender, it seems that many participants are plant (and animal) enthusiasts. For such groups, a public media campaign seems useful to attract participants. Platforms where people can search for projects according to their hobbies can also increase participation.
- For successful projects, there should be an easily accessible possibility for learning and to extend one’s knowledge about a topic. The importance for citizen scientists to be able to learn

new things has been reported in multiple studies (e.g. Hobbs and White, 2012; Johnson et al., 2014).

- People need to enjoy their participation. Thus, the activities need to be fun. This can possibly be enhanced with more options to participate (i.e., more choices and options to contribute).
- Social media elements are beneficial for online projects (Nov et al., 2014) to create social networks and allow people to comment on the contributions of others. This could help to form a community and ensure data quality (Serret et al., 2019). In Naturkalender, social interactions enable participants to help others and, therefore, provide teaching and learning experiences for the participants without requiring effort by the project administrators. This can be enhanced by giving users more competences (e.g. more rights for advanced users). This is in line with self-determination theory, according to which the ability to make competent actions and decisions autonomously and having the possibility to relate the project's topic to one's own interests leads to enhanced self-motivation (Ryan and Deci, 2000b).
- In this study, the super-users were in general older than the occasional participants. This is common for other projects as well (Sheppard et al., 2017; Wright et al., 2015). It might, therefore, be an effective strategy to focus recruitment on people above the age of 50. Once the habit is established, older people are more likely to contribute for extended periods (Sheppard et al., 2017; Venkatesh et al., 2012).
- Public platforms with available projects for interested people (e.g. scistarter.org) might be helpful for people who look for projects to participate. However, people are unlikely to search for an activity that they don't know, like observing water levels. Therefore, proactive and targeted social media marketing based on specific personal profiles and offline advertisements in local outdoor-based organisations, (e.g. bird-observers or dog-communities) or newspapers is still beneficial to reach a larger number of people.
- Respondents of the newer CrowdWater project were considerably more motivated to join by social pressure (*conformity*), i.e., because they were asked to help with the project. This might

be true for many projects in an early phase that still rely on family, friends or acquaintances to participate and promote the project. People who were motivated to join by a perceived social pressure may help a project in the beginning but later tend to contribute less or quit. We assume that Naturkalender participants were motivated more to join because of an interest in the project topic, in combination with a willingness and ability to share their expertise on the topic, which might indicate a perceived higher self-efficacy as defined by Phillips et al. (2018).

- The introduction of gamification elements increases the competitive element (Nov et al., 2014) and might attract new participants (Bowser et al., 2013a) but this might also decrease the intrinsic motivation of participants (Thiel and Fröhlich, 2017) or cause participants to make low-quality contributions in order to get more points (Bowser et al., 2013b). Thus, gamification should be applied cautiously and potential negative consequences should be evaluated beforehand. The respondents of this survey were not very motivated by competitive elements (low agreement for *achievement, face)*. Whether they did not like the existing leader board, or if it was not enough to trigger these motivations, remains to be investigated.

# 5 Conclusions

In this study, we used a questionnaire based on the citizen science literature to study the motivations that drive people to participate in citizen science projects and also reformulated the statements to investigate in how far their participation fulfilled these motivations. Participants of the CrowdWater and Naturkalender projects mainly joined the projects to contribute to science, to satisfy their interest in science and technology, to protect nature, contribute to the wellbeing of society, learn something new, and to be physically active. Not all of these initial motivations were fulfilled by participating in the projects. The respondents of both projects, for instance, agreed significantly less that their continued involvement was driven by a motivation to contribute to society (*universalism, societal concern*) and socialising with other people (*security and belongingness*) than they agreed on these aspects motivating them to join the projects in the first place. On the other hand, fun and enjoyment (*hedonism)* were not the primary motivations to become involved in the projects, but were essential motivators for continued participation. Roughly a third of all super-users (i.e., respondents contributing at least once a week) were 50-59 years old. This group of participants was most intrinsically motivated by enjoyment, learning and being physically active and outdoors, whereas participants in the other age groups valued the contribution they could make to science most. Respondents from the Naturkalender project were more motivated by enjoyment, learning (*self-direction*) and being outdoors and the physical activity (*stimulation*) than the CrowdWater respondents. Most of the fun and learning experience probably came from the social interaction and the information on plants and animals included in the Naturkalender app. Such a learning aspect was not available for CrowdWater, which probably explains why for CrowdWater respondents the primary motivation for continued participation were similar to the motivations for initial engagement: help with research (*universalism, research)*, protection of nature (*universalism, nature*) and acting according to their values and beliefs (*tradition*).

# 6 Acknowledgements

We thank all the respondents for filling in the questionnaire. This work would not have been possible without them. We, furthermore, thank Liat Levontin, Zohar Gilad and Shiraz Chako, for making the questionnaire available, and Assaf Shwartz, Liat Levontin and Zohar Gilad for hosting the Citizen Science Cost Action Workshop WG4 in March 2018, where the questionnaire was explained and discussed. We also thank Philipp Hummer of SPOTTERON for facilitating the communication with the project leaders, of Naturkalender Thomas Hübner of the Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Karin Schroll and Isabella Ostovary of Lacon – Landschaftsplanung Consultng for the provision of Naturkalender data and the collaboration, and Florian Heigl and Didone Frigerio of the projects RoadKill and Forschen im Almtal for sending the questionnaire to their participants. This study was funded by the Swiss National Science Foundation (project 163008, CrowdWater).

# 7 References

Alender, B., 2016. Understanding volunteer motivations to particpate in citizen science projects: a deeper look at water quality monitoring. J. Sci. Commun. 15, 2–19.

Bandura, A., 1997. Self Efficacy - The Exercise of Control, 1st ed. Macmillan International Higher Education.

Batson, C.D., Ahmad, N., Tsang, J.-A., 2002. Four Motives for Community Involvement. J. Soc. Issues 58, 429–445. https://doi.org/10.1111/1540-4560.00269

Beaubien, E.G., Hamann, A., 2011. Plant phenology networks of citizen scientists: recommendations from two decades of experience in Canada. Int. J. Biometeorol. 55, 833–841. https://doi.org/10.1007/s00484-011-0457-y

Bell, S., Marzano, M., Cent, J., Kobierska, H., Podjed, D., Vandzinskaite, D., Reinert, H., Armaitiene, A., Grodzińska-Jurczak, M., Muršič, R., 2008. What counts? Volunteers and their organisations in the recording and monitoring of biodiversity. Biodivers. Conserv. 17, 3443–3454. https://doi.org/10.1007/s10531-008-9357-9

Beza, E., Steinke, J., van Etten, J., Reidsma, P., Fadda, C., Mittra, S., Mathur, P., Kooistra, L., 2017. What are the prospects for citizen science in agriculture? Evidence from three continents on motivation and mobile telephone use of resource-poor farmers. PLoS One 12, e0175700. https://doi.org/10.1371/journal.pone.0175700

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., Preece, J., 2013a. Using gamification to inspire new citizen science volunteers, in: Proceedings of the First International Conference on Gameful Design, Research, and Applications - Gamification ’13. Stratford, Ontario, Canada, pp.

18–25. https://doi.org/10.1145/2583008.2583011

Bowser, A., Hansen, D., Preece, J., 2013b. Gamifying Citizen Science: Lessons and Future Directions. Work. Des. Gamification Creat. Gameful Play. Exp.

Clary, E.G., Snyder, M., 1999. The Motivations to Volunteer. Curr. Dir. Psychol. Sci. 8, 156–159. https://doi.org/10.1111/1467-8721.00037

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334. https://doi.org/10.1007/BF02310555

Curtis, V., 2015. Motivation to Participate in an Online Citizen Science Game. Sci. Commun. 37, 723–746. https://doi.org/10.1177/1075547015609322

de Vries, M., Land-Zandstra, A., Smeets, I., 2019. Citizen Scientists' Preferences for Communication of Scientific Output: A Literature Review. Citiz. Sci. Theory Pract. 4, 1–13. https://doi.org/10.5334/cstp.136

Deci, E.L., Ryan, R.M., 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. Psychol. Inq. 11, 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Domroese, M.C., Johnson, E.A., 2017. Why watch bees? Motivations of citizen science volunteers in the Great Pollinator Project. Biol. Conserv. 208, 40–47. https://doi.org/10.1016/j.biocon.2016.08.020

Eisinga, R., Grotenhuis, M. Te, Pelzer, B., 2013. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int. J. Public Health 58, 637–642. https://doi.org/10.1007/s00038-012-0416-3

Finkelstien, M.A., 2009. Intrinsic vs. extrinsic motivational orientations and the volunteer process. Pers. Individ. Dif. 46, 653–658. https://doi.org/10.1016/j.paid.2009.01.010

Frensley, T., Crall, A., Stern, M., Jordan, R., Gray, S., Prysby, M., Newman, G., Hmelo-Silver, C., Mellor, D., Huang, J., 2017. Bridging the Benefits of Online and Community Supported Citizen Science: A Case Study on Motivation and Retention with Conservation-Oriented Volunteers. Citiz. Sci. Theory Pract. 2, 4. https://doi.org/10.5334/cstp.84

Fuccillo, K.K., Crimmins, T.M., de Rivera, C.E., Elder, T.S., 2015. Assessing accuracy in citizen science-based plant phenology monitoring. Int. J. Biometeorol. 59, 917–926. https://doi.org/10.1007/s00484-014-0892-7

Füchslin, T., Schäfer, M.S., Metag, J., 2019. Who wants to be a citizen scientist? Identifying the potential of citizen science and target segments in Switzerland. Public Underst. Sci. 28, 652–668. https://doi.org/10.1177/0963662519852020

Furnham, A., 1986. Response bias, social desirability and dissimulation. Pers. Individ. Dif. 7, 385–400. https://doi.org/10.1016/0191-8869(86)90014-0

Geoghegan, H., Dyke, A., Pateman, R., West, S., Everett, G., 2016. Understanding Motivations for Citizen Science. Final Report on behalf of the UK Environmental Observation Framework (UKEOF).

George, D., Mallery, P., 2003. SPSS for Windows step by step: A simple guide andreference., 4th ed. Allyn & Bacon, Boston.

Haklay, M., 2018. Participatory citizen science, Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, UK. https://doi.org/10.14324/111.9781787352339

Hobbs, S.J., White, P.C.L., 2012. Motivations and barriers in relation to community participation in biodiversity recording. J. Nat. Conserv. 20, 364–373. https://doi.org/10.1016/j.jnc.2012.08.002

Irwin, A., 2018. No PhDs needed: how citizen science is transforming research. Nature 562, 480–482. https://doi.org/10.1038/d41586-018-07106-5

Israel, G.D., 1992. Determination of sample size. Progr. Eval. Organ. Dev. Fact Sheet PEOD-6.

Johnson, M.F., Hannah, C., Acton, L., Popovici, R., Karanth, K.K., Weinthal, E., 2014. Network environmentalism: Citizen scientists as agents for environmental advocacy. Glob. Environ. Chang. 29, 235–245. https://doi.org/10.1016/j.gloenvcha.2014.10.006

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., Puntenney-Desmond, K., Seibert, J., van Meerveld, I., 2018. Testing the waters: Mobile apps for crowdsourced streamflow data. Eos (Washington. DC). 99. https://doi.org/10.1029/2018EO096355

Land-Zandstra, A.M., Devilee, J.L.A., Snik, F., Buurmeijer, F., van den Broek, J.M., 2016a. Citizen science on a smartphone: Participants' motivations and learning. Public Underst. Sci. 25, 45–60. https://doi.org/10.1177/0963662515602406

Land-Zandstra, A.M., van Beusekom, M.M., Koppeschaar, C.E., van den Broek, J.M., 2016b. Motivation and learning impact of Dutch flu-trackers. J. Sci. Commun. 15, 1–26.

Levontin, L., Gilad, Z., Chako, S., 2018. Questionare for the Motivation for Citizen Science Scale.

Locke, M., Ellis, A., Davis Smith, J., 2003. Hold on to what you've got: the volunteer retention literature. Volunt. Action 5, 81–99.

Martin, A.J., 2007. Examining a multidimensional model of student motivation and engagement using a construct validation approach. Br. J. Educ. Psychol. 77, 413–440. https://doi.org/10.1348/000709906X118036

Nederhof, A.J., 1985. Methods of coping with social desirability bias: A review. Eur. J. Soc. Psychol. 15, 263–280. https://doi.org/10.1002/ejsp.2420150303

Nov, O., Arazy, O., Anderson, D., 2014. Scientists@Home: What drives the quantity and quality of online citizen science participation? PLoS One 9, 1–11. https://doi.org/10.1371/journal.pone.0090375

Pandya, R., Dibner, K.A., 2018. Learning Through Citizen Science, Learning Through Citizen Science: Enhancing Opportunities by Design. National Academies Press, Washington, D.C.

https://doi.org/10.17226/25183

Parrish, J.K., Burgess, H., Weltzin, J.F., Fortson, L., Wiggins, A., Simmons, B., 2018. Exposing the Science in Citizen Science: Fitness to Purpose and Intentional Design. Integr. Comp. Biol. 1–11. https://doi.org/10.1093/icb/icy032

Penner, L.A., 2002. Dispositional and Organizational Influences on Sustained Volunteerism: An Interactionist Perspective. J. Soc. Issues 58, 447–467. https://doi.org/10.1111/1540-4560.00270

Phillips, T., Porticella, N., Constas, M., Bonney, R., 2018. A Framework for Articulating and Measuring Individual Learning Outcomes from Participation in Citizen Science. Citiz. Sci. Theory Pract. 3, 3. https://doi.org/10.5334/cstp.126

Phillips, T.B., Ballard, H.L., Lewenstein, B. V., Bonney, R., 2019. Engagement in science through citizen science: Moving beyond data collection. Sci. Educ. 103, 665–690. https://doi.org/10.1002/sce.21501

Pieper, K.J., Martin, R., Tang, M., Walters, L., Parks, J., Roy, S., Devine, C., Edwards, M.A., 2018. Evaluating Water Lead Levels During the Flint Water Crisis. Environ. Sci. Technol. 52, 8124–8132. https://doi.org/10.1021/acs.est.8b00791

Raddick, J.M., Bracey, G., Gay, P.L., Lintott, C.J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J., 2013. Galaxy Zoo: Motivations of Citizen Scientists. Astron. Educ. Rev. 12. https://doi.org/10.3847/AER2011021

Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C.J., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J., 2010. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. Astron. Educ. Rev. 9. https://doi.org/10.3847/AER2009036

Rey-Mazón, P., Keysar, H., Dosemagen, S., D'Ignazio, C., Blair, D., 2018. Public Lab: Community-Based Approaches to Urban and Environmental Health and Justice. Sci. Eng. Ethics 24, 971–997. https://doi.org/10.1007/s11948-018-0059-8

Rinderer, M., Kollegger, A., Fischer, B.M.C., Stähli, M., Seibert, J., 2012. Sensing with boots and trousers - qualitative field observations of shallow soil moisture patterns. Hydrol. Process. 26, 4112–4120. https://doi.org/10.1002/hyp.9531

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., Jacobs, D., 2012. Dynamic changes in motivation in collaborative citizen-science projects, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW ’12. ACM Press, New York, New York, USA, p. 217. https://doi.org/10.1145/2145204.2145238

Roy, H.E., Pocock, M.J.O., Preston, C.D., Roy, D.B., Savage, J., Tweddle, J.C., Robinson, L.D., 2012. Understanding Citizen Science and Environmental Monitoring. Understanding Citizen Scie nce & Environmental Monitoring. Final Report on behalf of UK-EOF. NERC Centre for Ecology & Hydrology and Natural History Museum.

Ryan, R.L., Kaplan, R., Grese, R.E., 2001. Predicting Volunteer Commitment in Environmental Stewardship Programmes. J. Environ. Plan. Manag. 44, 629–648. https://doi.org/10.1080/09640560120079948

Ryan, R.M., Deci, E.L., 2000a. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemp. Educ. Psychol. 25, 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R.M., Deci, E.L., 2000b. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. 55, 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Schwartz, S.H., 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. pp. 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6

Schwartz, S.H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., Konty, M., 2012. Refining the theory of basic individual values. J. Pers. Soc. Psychol. 103, 663–688. https://doi.org/10.1037/a0029393

Seibert, J., Strobl, B., Etter, S., Hummer, P., van Meerveld, H.J. (Ilja), 2019a. Virtual Staff Gauges for Crowd-Based Stream Level Observations. Front. Earth Sci. 7. https://doi.org/10.3389/feart.2019.00070

Seibert, J., van Meerveld, H.J., Etter, S., Strobl, B., Assendelft, R., Hummer, P., 2019b. Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen? Hydrol. und Wasserbewirtschaftung 63. https://doi.org/10.5675/HyWa_2019.2_1

Serret, H., Deguines, N., Jang, Y., Lois, G., Julliard, R., 2019. Data Quality and Participant Engagement in Citizen Science: Comparing Two Approaches for Monitoring Pollinators in France and South Korea. Citiz. Sci. Theory Pract. 4, 22. https://doi.org/10.5334/cstp.200

Sheppard, S.A., Turner, J., Thebault-Spieker, J., Zhu, H., Terveen, L., 2017. Never Too Old, Cold or Dry to Watch the Sky. Proc. ACM Human-Computer Interact. 1, 1–21. https://doi.org/10.1145/3134729

Szanton, S.L., Walker, R.K., Roberts, L., Thorpe, R.J., Wolff, J., Agree, E., Roth, D.L., Gitlin, L.N., Seplaki, C., 2015. Older adults' favorite activities are resoundingly active: Findings from the NHATS study. Geriatr. Nurs. (Minneap). https://doi.org/10.1016/j.gerinurse.2014.12.008

Thiel, S.-K., Fröhlich, P., 2017. Progress in Location-Based Services 2016, Progress in Location-Based Services 2016, Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-47289-8

Thornhill, I., Loiselle, S., Clymans, W., van Noordwijk, C.G.E., 2019. How citizen scientists can enrich freshwater science as contributors, collaborators, and co-creators. Freshw. Sci. 38, 231–235. https://doi.org/10.1086/703378

Venkatesh, Thong, Xu, 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Q. 36, 157. https://doi.org/10.2307/41410412

West, S., Pateman, R., 2016. Recruiting and Retaining Participants in Citizen Science: What Can Be Learned from the Volunteering Literature? Citiz. Sci. Theory Pract. 1, 1–10. https://doi.org/10.5334/cstp.8

Wright, D.R., Underhill, L.G., Keene, M., Knight, A.T., 2015. Understanding the Motivations and Satisfactions of Volunteers to Improve the Effectiveness of Citizen Science Programs. Soc. Nat. Resour. 28, 1013–1029. https://doi.org/10.1080/08941920.2015.1054976

# 8 Supplemental Material

*Table S1 Categories according to Batson et al. (2002) and the corresponding statements from the questionnaire of* Levontin et al. (2018). *The statements for the fulfilment were adapted from those in the engagement part.*

| Category | Potential Motivations for Engagement | Potentially Fulfilled Motivational goals |
|---|---|---|
| **Altruism** | I want to make scientific knowledge accessible to the public | By contributing to this project I can make scientific knowledge accessible to the public |
| | I do this activity because I am happy to help | By doing this activity I can help others |
| **Collectivism** | It's a nice family activity | By contributing to this project I get to have some good times with my family |
| | I want to contribute to the future of humanity | By contributing to this project I can contribute to the future of humanity |
| | I want to make the world a better place | By contributing to this project I can make the world a better place |
| | It's a teaching opportunity | Participating in this project provided me a teaching opportunity |
| | I want to contribute to science | This activity helped me to contribute to science |
| | I want to contribute to the knowledge about this topic | By contributing to this project I can contribute to the knowledge about this topic |
| **Egoism extrinsic** | Volunteering makes me feel important | Volunteering in this project makes me feel important |
| | Other people I know are participating | - |
| | Other people think positively about my contribution to this project | - |
| | I am seeking fame | I can satisfy my need for fame by doing this activity |
| | I was requested to participate by somebody | - |
| | I want to be part of this volunteers' community | - |
| | I want to receive recognition | I can get recognition for participating in this project |
| | I want to socialize with other people | This project is an opportunity to socialize with other people |
| | I want to enhance my reputation | This activity helps me to enhance my reputation |
| | I expect something in return | - |
| | I want to meet people with similar interests | By participating in this project, I meet people with similar interests |

| | | |
|---|---|---|
| | I want to gain social status | This activity increased my social status |
| | I like to compete with others | I can compete with others in this project |
| **Egoism intrinsic** | I want to spend time in nature | By participating in this project I get to spend more time in nature |
| | I am interested in the topic of this project | - |
| | I want to learn new skills or new knowledge | This activity taught me new skills or knowledge |
| | I am interested in science and technology. | This activity satisfies my interest in science and technology |
| | This activity is related to another hobby I have R | - |
| | I want to have fun | This activity is fun for me |
| | - | I enjoy this activity |
| | I want to do something meaningful | This activity is meaningful |
| | I want to do some physical activity | By participating in this project I am physically active |
| | I want to share my knowledge and my experience | By contributing to this project I can share my knowledge and experiences |
| | I want to spend more time outdoors | By participating in this project I get to spend more time outdoors |
| | I strive to challenge myself | This activity challenged myself |
| **Principlism** | My beliefs and/or my values motivated me to participate. | Helping with this project is according to my beliefs and/or my values |
| | I want to contribute to conservation | By contributing to this project I can contribute to conservation |
| | I want to raise public awareness of this topic | By contributing to this project I can raise public awareness of this topic |

*Table S2 Categories according to Schwartz et al. (2012) and the corresponding statements from the questionnaire of* Levontin et al. (2018).

| **Categories** | **Conceptual definition*** | **Potential Motivations for Engagement** | **Potentially Fulfilled Motivational goals** |
|---|---|---|---|
| Achievement | Personal success through demonstrating competence according to social standards<br>Achieving goals according to social standards and thereby demonstrating competence. | I am seeking fame<br>I want to do something meaningful<br>I like to compete with others | This activity is meaningful<br>I can compete with others in this project<br>I can satisfy my need for fame by doing this activity |
| Benevolence, caring | Preservation and enhancement of the welfare of people with whom one is in frequent personal contact<br>Improving or preserving the wellbeing of people that are relevant in one's everyday life. | It's a nice family activity<br>I do this activity because I am happy to help | By doing this activity I can help others<br>By contributing to this project I get to have some good times with my family |
| Conformity | Trying to act in a way that does not harm or upset anyone and fulfils social expectations or norms | Other people I know are participating<br>I was requested to participate by somebody | - |
| Face | Security and power by avoiding humiliation and maintaining a good reputation. | Other people think positively about my contribution to this project<br>I want to enhance my reputation | This activity helps me to enhance my reputation |
| Hedonism | Experience pleasure and enjoyment physically or mentally | I want to have fun | I enjoy this activity<br>This activity is fun for me |
| Power, Dominance | Maintaining social status and prestige by controlling and dominating other people. | Volunteering makes me feel important<br>I want to receive recognition<br>I want to gain social status | I can get recognition for participating in this project<br>Volunteering in this project makes me feel important<br>This activity increased my social status |
| Power, resources | Maintaining or achieving social status and prestige by controlling or acquiring resources | I expect something in return | - |

| | | | |
|---|---|---|---|
| Security and belongingness | Safety by feeling connected to a community | I want to be part of this volunteers' community<br>I want to socialize with other people<br>I want to meet people with similar interests | By participating in this project, I meet people with similar interests<br>This project is an opportunity to socialize with other people |
| Self-direction | Independent exploring, learning and being creative | I want to learn new skills or new knowledge<br>I am interested in the topic of this project<br>I am interested in science and technology. | This activity satisfies my interest in science and technology<br>This activity taught me new skills or knowledge |
| Stimulation and routine break | Doing exciting, and new things that might also challenge oneself | This activity is related to another hobby I have R<br>I strive to challenge myself | This activity challenged myself |
| Stimulation, being outside and active | Doing exciting, and new things that might challenge oneself in the outdoors and being physically active. | I want to spend time in nature<br>I want to do some physical activity<br>I want to spend more time outdoors | By participating in this project I am physically active<br>By participating in this project I get to spend more time in nature<br>By participating in this project I get to spend more time outdoors |
| Tradition | Upholding traditional principles, values, and customs of a culture or religion | My beliefs and/or my values motivated me to participate. | Helping with this project is according to my beliefs and/or my values |
| Universalism, help with research | Upholding the value of science and support it. | I want to contribute to science<br>I want to contribute to the knowledge about this topic | By contributing to this project I can contribute to the knowledge about this topic<br>This activity helped me to contribute to science |

| | | | |
|---|---|---|---|
| Universalism, nature | Upholding the value of nature and protecting it. | I want to contribute to conservation<br>I want to raise public awareness of this topic | By contributing to this project I can raise public awareness of this topic<br>By contributing to this project I can contribute to conservation |
| Universalism, societal concern | Appreciating the value of society, protect and improve it | I want to contribute to the future of humanity<br>I want to make scientific knowledge accessible to the public<br>I want to make the world a better place | By contributing to this project I can contribute to the future of humanity<br>By contributing to this project I can make the world a better place<br>By contributing to this project I can make scientific knowledge accessible to the public |
| Universalism, teaching | Upholding the value of teaching and sharing experiences. | It's a teaching opportunity<br>I want to share my knowledge and my experience | By contributing to this project I can share my knowledge and experiences<br>Participating in this project provided me a teaching opportunity |

***** adapted from Schwartz et al. (2012)

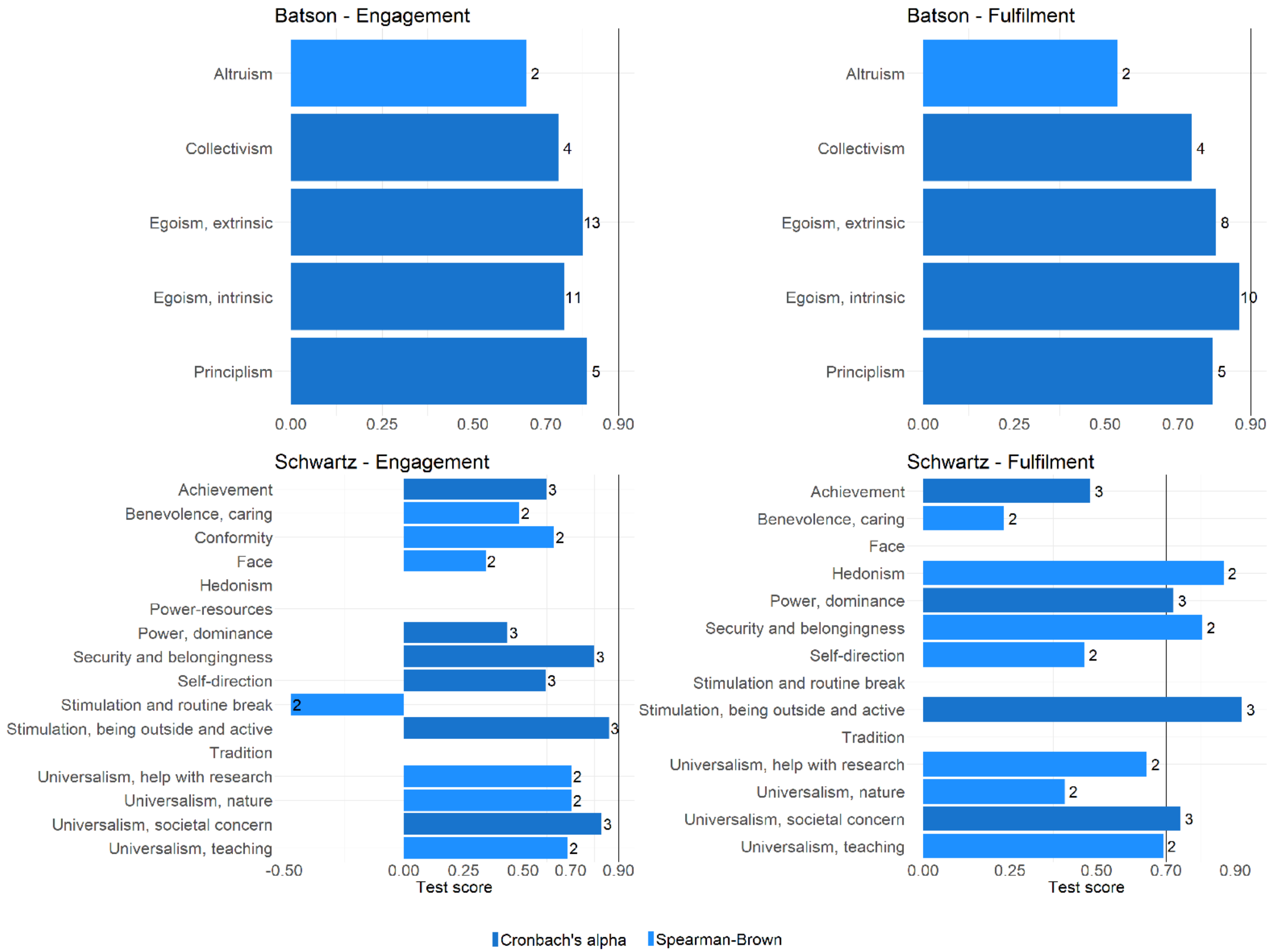



*Figure S1 Results of the reliability analysis of the categories and the corresponding statements using Cronbach's alpha (Cronbach, 1951) in dark blue and Spearman-Brown (Eisinga et al., 2013) in light blue for the categories with two statements. A score could not be calculated for the categories that consist of only one item (no bars). The numbers at the end of the bars indicate the number of statements in the category.*

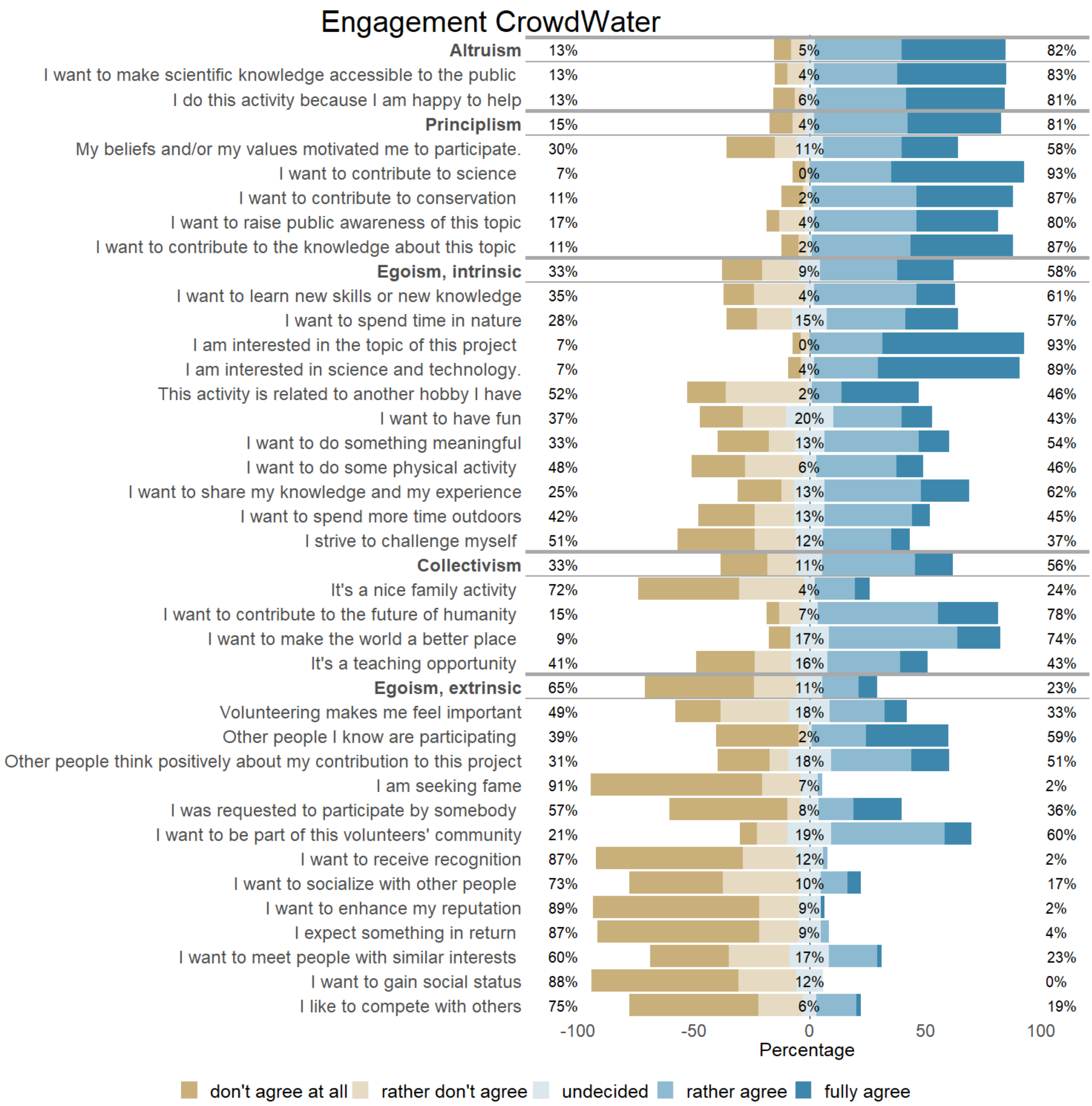
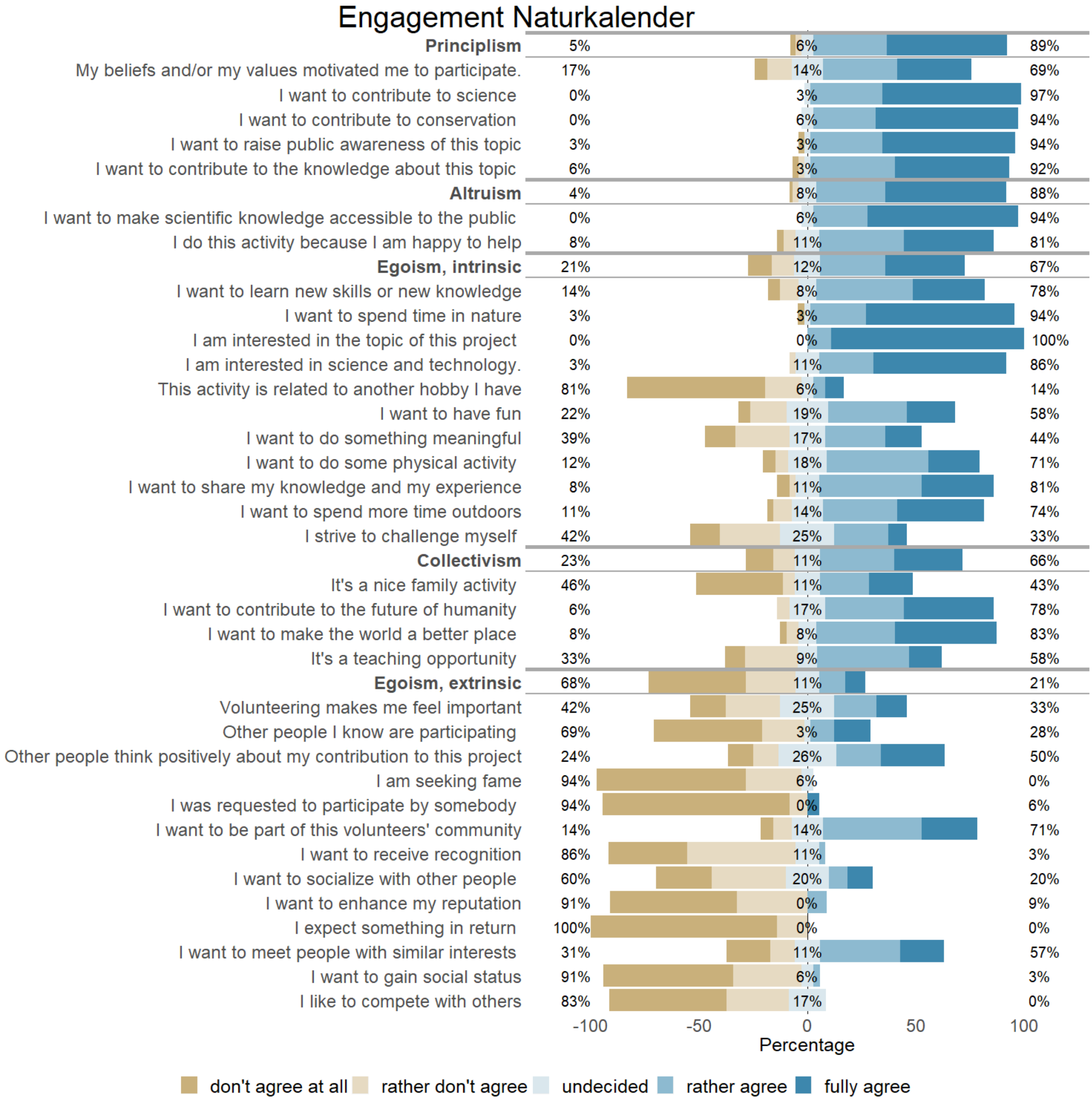



*Figure S2 The agreement to the statements for initial engagement for the CrowdWater project grouped per category of Batson et al. (2002) (in bold font).*




930

931 *Figure S3 The agreement to the statements for initial engagement for the Naturkalender project grouped per*
932 *category of the Batson-scheme (shown in bold).*

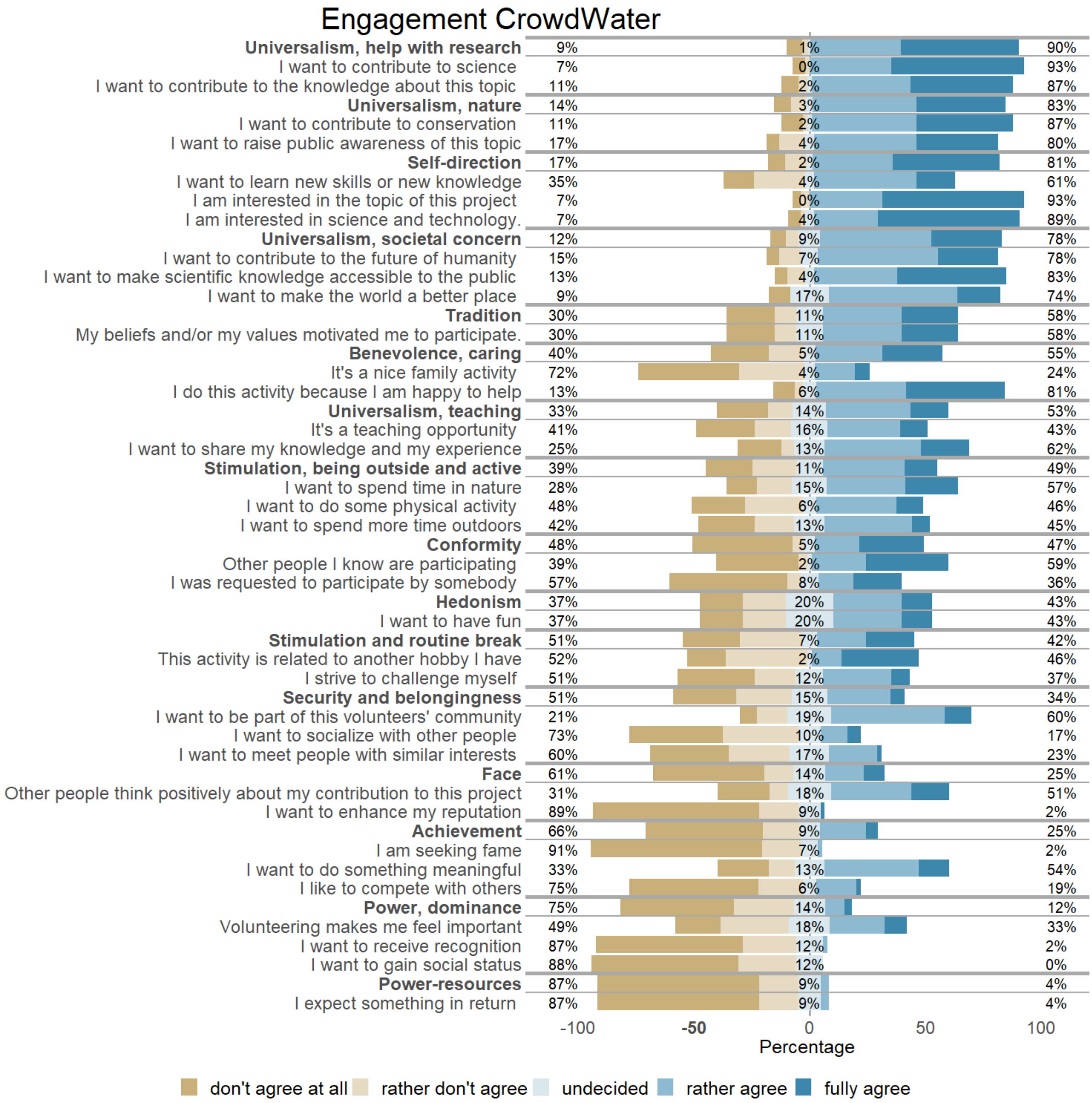
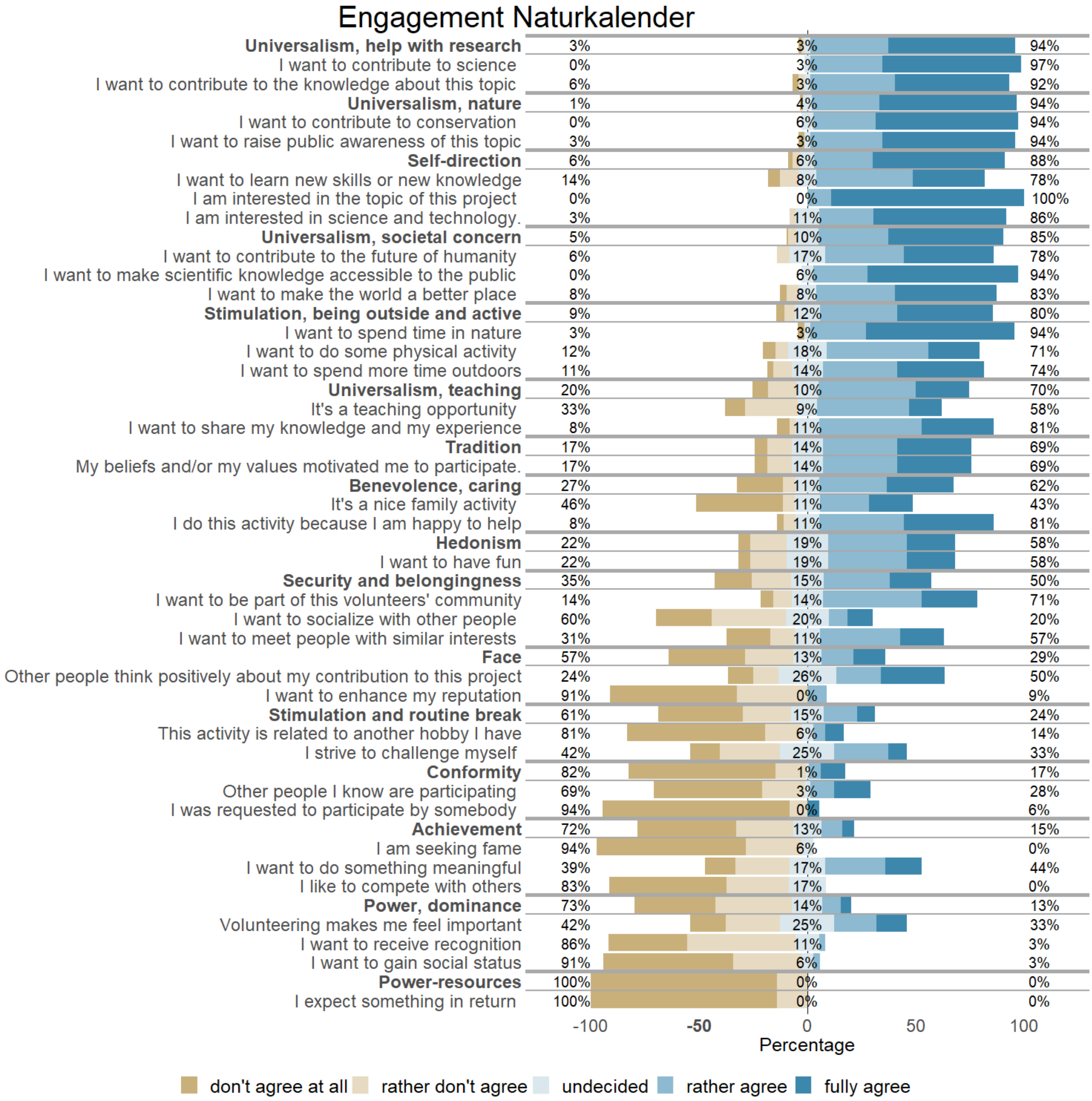



*Figure S4 The agreement to the statements for initial engagement for the CrowdWater project grouped per category of the Schwartz-scheme (in bold font).*




936

937 *Figure S5 The agreement to the statements for initial engagement for the Naturkalender project of the Schwartz-*
938 *scheme.*

939

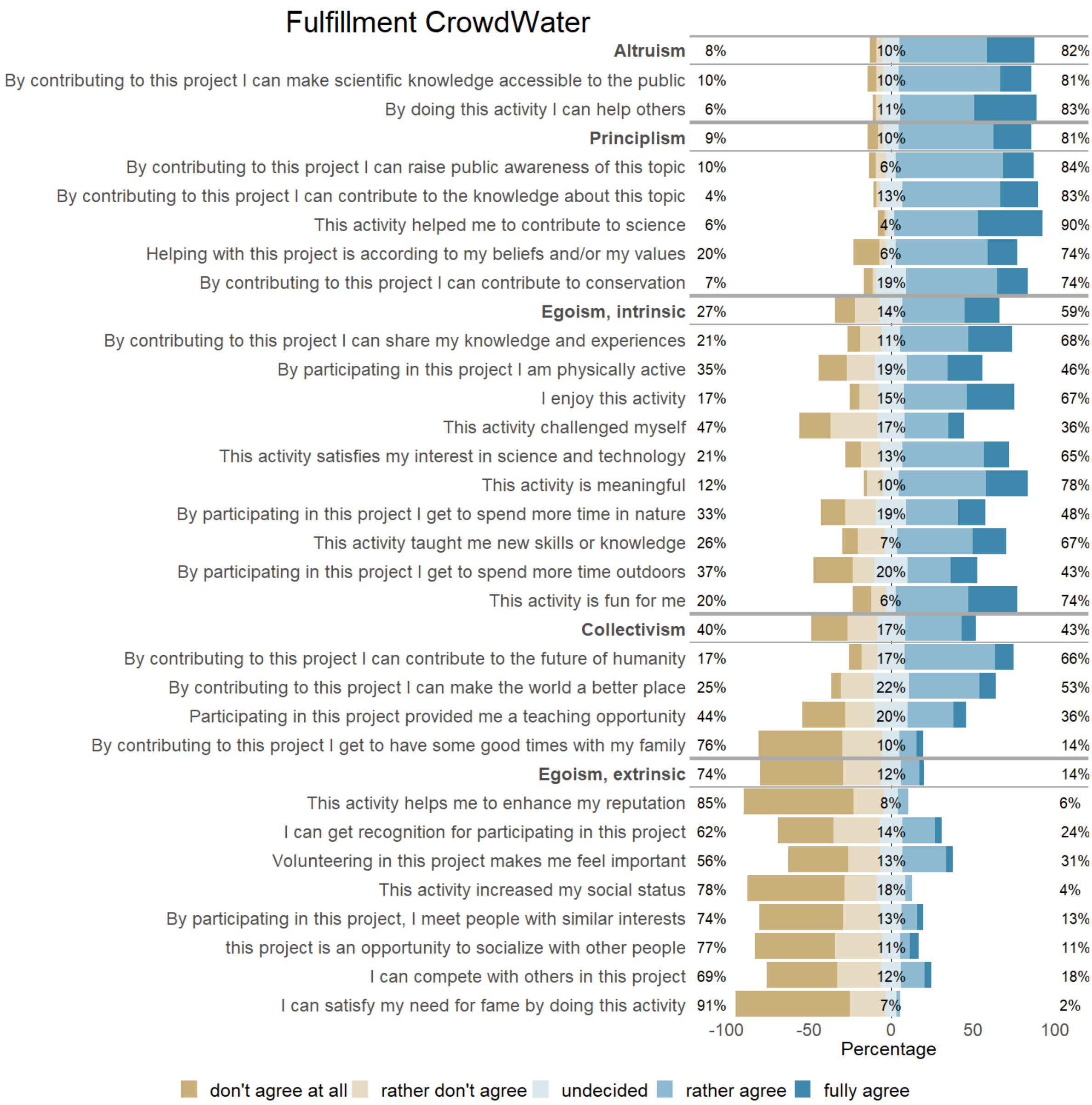



*Figure S6 Agreement of CrowdWater participants to the statements related to how their initial motivations were fulfilled by participation in the in the project grouped per category of the Batson-scheme.*

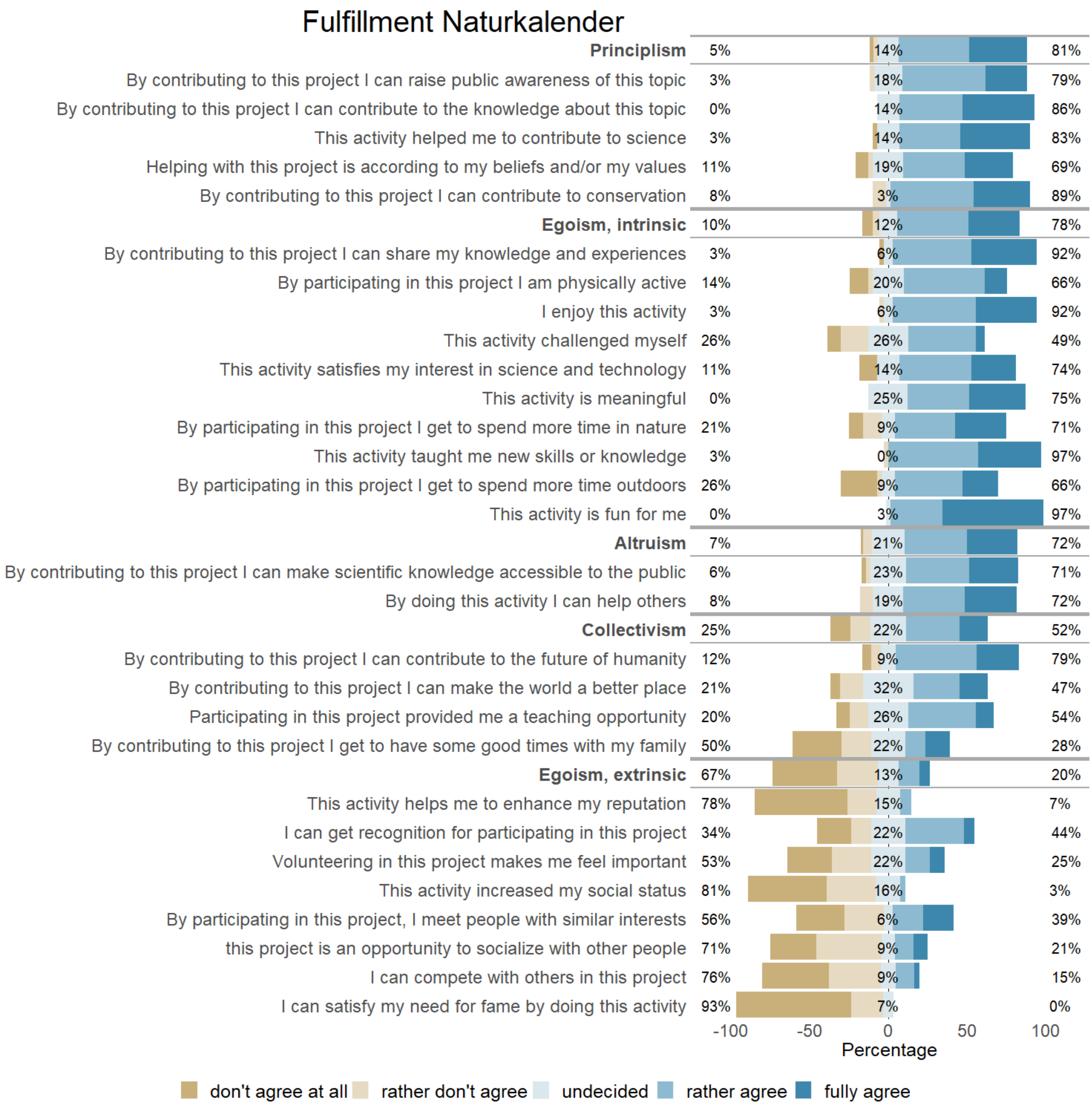



*Figure S7 Agreement of Naturkalender participants to the statements related to how their initial motivations were fulfilled by participation in the in the project grouped per category of the Batson-scheme.*

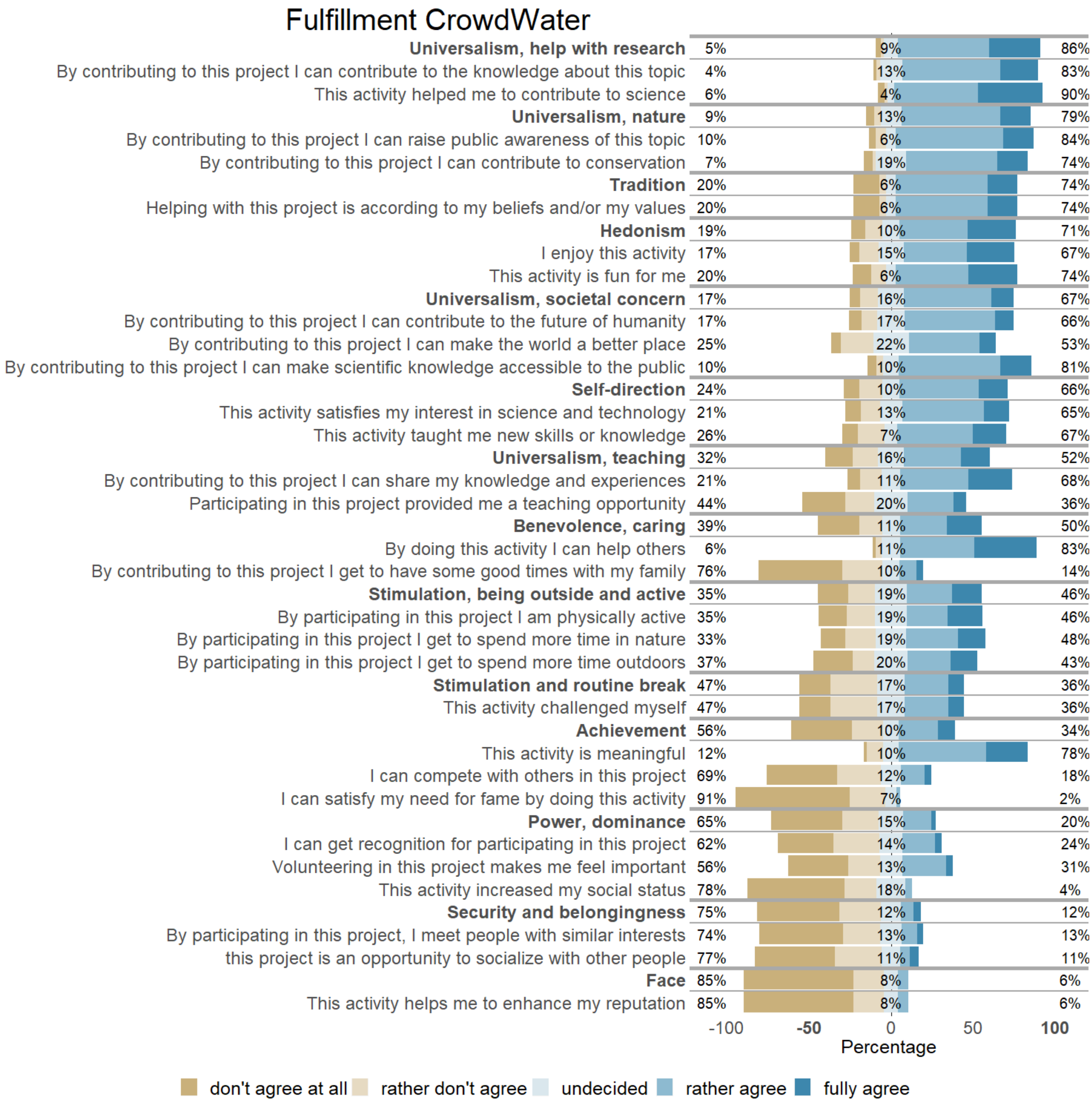



*Figure S8 Agreement of CrowdWater participants to the statements related to how their initial motivations were fulfilled by participation in the in the project grouped per category of the Schwartz-scheme.*

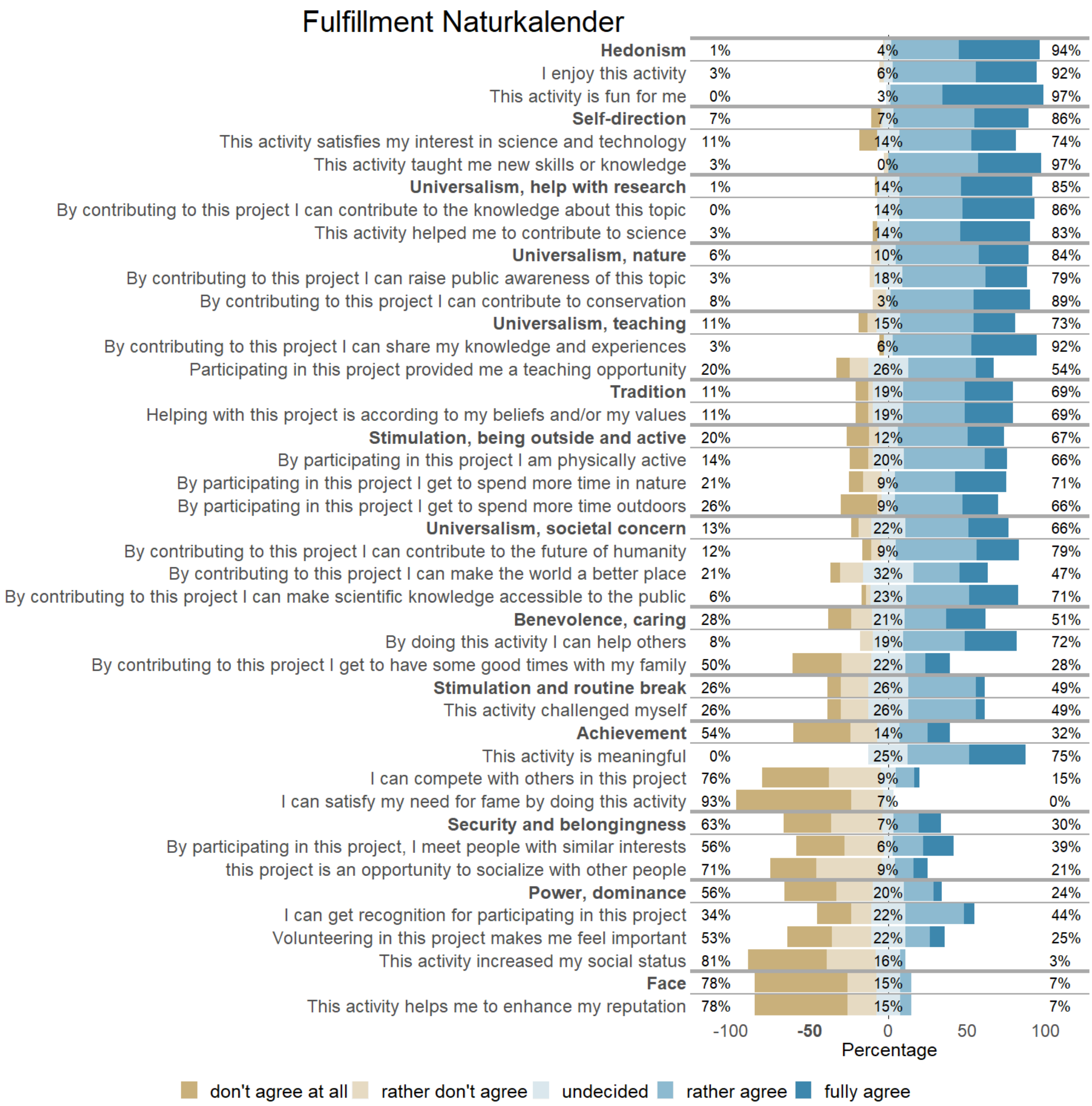



*Figure S9 Agreement of Naturkalender participants to the statements related to how their initial motivations were fulfilled by participation in the in the project grouped per category of the Schwartz-scheme.*

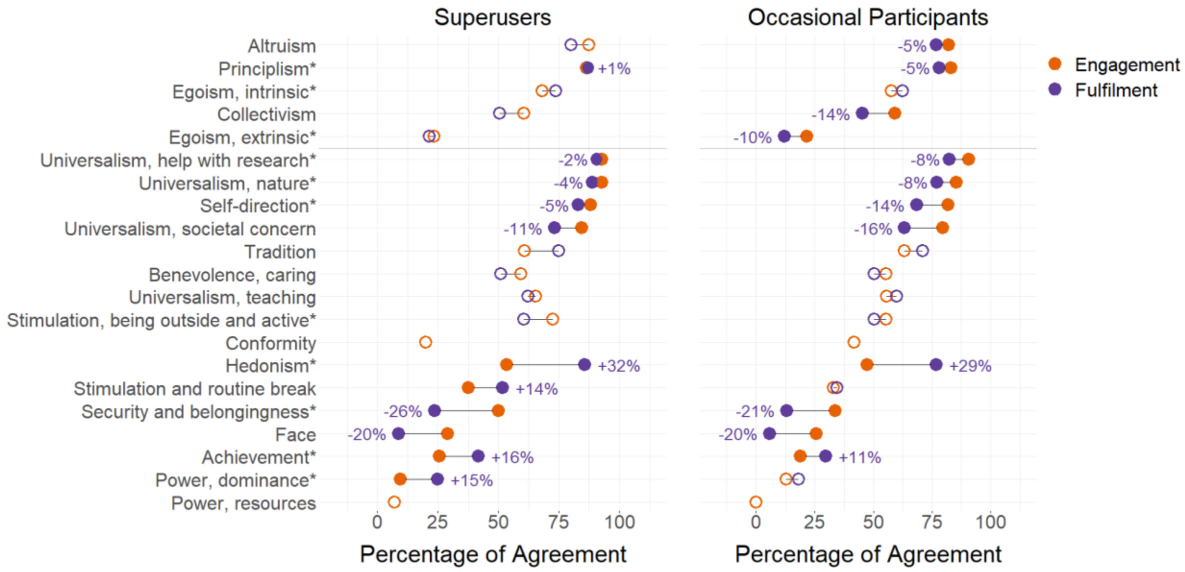



*Figure S10 Comparison of the percentage of super-users and occasional participants who agree to the different statements on motivations for engagement (orange) and whether these are fulfilled by participating in the project (purple). Significant differences in the median agreement for engagement and fulfilment are shown in solid circles; insignificant differences by open circles The graph elements are sorted by decreasing agreement to the categories in the engagement part in CrowdWater project to enable comparison with Figure 4. The asterisks in the y-axis labels indicates a significant difference between the super-users and the occasional participants.*

Citizen Science Motivation EN

Thank you very much for your time!

**You are invited to participate in this survey because you are part of at least one of the citizen science projects on the SPOTTERON-platform (www.spotteron.net).**

**I will distribute 10 "Hydrosommelier-Bottles" randomly between all participants who completed the survey. In the end of the survey you are asked to enter your e-mail adress in case you want to to sign up for the list of potential recipients.**

**This survey is part of my PhD thesis at the University of Zurich, Switzerland. I work in the CrowdWater project and am trying to find out more about what motivates people to participate in citizen science projects. In this survey I will ask you some questions in the form of statements. Please answer them on a scale from “don’t agree at all” to “fully agree”, depending on how well the statements apply to you.**

**In the first part (questions 5-11) you are asked about the reasons you chose to participate in one or multiple citizen science projects on the SPOTTERON platform. If you participated in more than one project, please choose the one to which you contributed first. Please don’t consider if your expectations have been fulfilled in the first part.**

**In the second part (questions 12-17) you are asked whether or not you agree to statements about how well the expectations have been met for participating in the project in the first place. While doing so, please consider how you feel about participating in the citizen sciene project today.**

**It would be very helpful if you could answer four short questions about yourself at the end.**

**Your answers will be recorded and stored anonymously. It will take about 10 minutes to fill in the survey.**

**Thank you very much for your help!**
**Simon Etter**
**(simon.etter@geo.uzh.ch)**

* 1. Please create your personal code to enable the comparison with a potential follow-up study. The code consists of the first two letters of your mothers first name, the first two letters of your fathers first name and the numbers of the day of your birth date.

Example:

Mothers First Name: **An**ita

Fathers First Name: **Ro**bert

Birthday: **01**.02.1980

Example personal Code: **AnRo01**

____________________

* 2. In which of the SPOTTERON-based projects do or did you participate? (Multiple answers are possible)?

- [ ] CrowdWater
- [ ] Naturkalender ZAMG
- [ ] Naturkalender NÖ
- [ ] Naturkalender Steiermark
- [ ] Roadkill
- [ ] Forschen im Almtal
- [ ] GLOBAL2000 Naturputzer
- [ ] StreetArt
- [ ] Waldrapp
- [ ] Was geht ab?
- [ ] Fågelbär
- [ ] GEFABE
- [ ] Fjällkalendern
- [ ] other project

3. When did you join the project?

- ○ Days ago
- ○ Weeks ago
- ○ Months ago
- ○ in the second half of 2017
- ○ in the first half of 2017
- ○ in 2016
- ○ in 2015
- ○ before 2015

4. How often do you contribute to the project?

- ○ more than once a day
- ○ once a day
- ○ every few days
- ○ weekly
- ○ every few weeks
- ○ monthly
- ○ less than monthly
- ○ I contributed only a few times (1-3 times)
- ○ I have never contributed

Citizen Science Motivation EN

## What made you decide to participate in a citizen science project?

* 5. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to learn new skills or new knowledge | ○ | ○ | ○ | ○ | ○ | ○ |
| Volunteering makes me feel important | ○ | ○ | ○ | ○ | ○ | ○ |
| Other people I know are participating | ○ | ○ | ○ | ○ | ○ | ○ |
| It's a nice family activity | ○ | ○ | ○ | ○ | ○ | ○ |
| Other people think positively about my contributions to this project | ○ | ○ | ○ | ○ | ○ | ○ |

* 6. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to contribute to the future of humanity | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to spend time in nature | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to make scientific knowledge accessible to the public | ○ | ○ | ○ | ○ | ○ | ○ |
| I am seeking fame | ○ | ○ | ○ | ○ | ○ | ○ |
| I am interested in the topic of this project | ○ | ○ | ○ | ○ | ○ | ○ |

* 7. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I am interested in science and technology. | ○ | ○ | ○ | ○ | ○ | ○ |
| I was requested to participate by somebody | ○ | ○ | ○ | ○ | ○ | ○ |
| This activity is related to another hobby I have | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to make the world a better place | ○ | ○ | ○ | ○ | ○ | ○ |
| It's a teaching opportunity | ○ | ○ | ○ | ○ | ○ | ○ |

* 8. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to have fun | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to be part of this volunteers' community | ○ | ○ | ○ | ○ | ○ | ○ |
| My beliefs and/or my values motivated me to participate. | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to receive recognition | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to do something meaningful | ○ | ○ | ○ | ○ | ○ | ○ |

* 9. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to contribute to science | ○ | ○ | ○ | ○ | ○ | ○ |
| I do this activity because I am happy to help | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to do some physical activity | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to socialize with other people | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to share my knowledge and my experience | ○ | ○ | ○ | ○ | ○ | ○ |

* 10. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to spend more time outdoors | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I want to contribute to conservation | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I strive to challenge myself | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I want to enhance my reputation | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I expect something in return | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 11. Why did you join the citizen science project?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I want to meet people with similar interests | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I want to raise public awareness of this topic | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I want to gain social status | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I like to compete with others | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I want to contribute to the knowledge about this topic | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

Citizen Science Motivation EN

## How have your expectations been fulfilled by the participation in the project?

* 12. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| By contributing to this project I can raise public awareness of this topic | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By contributing to this project I can contribute to the knowledge about this topic | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By contributing to this project I can share my knowledge and experiences | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By participating in this project I am physically active | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| This activity helped me to contribute to science | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 13. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| This activity helps me to enhance my reputation | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By contributing to this project I can contribute to the future of humanity | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I can get recognition for participating in this project | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Volunteering in this project makes me feel important | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I enjoy this activity | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 14. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| By contributing to this project I can make the world a better place | ○ | ○ | ○ | ○ | ○ | ○ |
| Helping with this project is according to my beliefs and/or my values | ○ | ○ | ○ | ○ | ○ | ○ |
| This activity challenged myself | ○ | ○ | ○ | ○ | ○ | ○ |
| This activity increased my social status. | ○ | ○ | ○ | ○ | ○ | ○ |
| This activity satisfies my interest in science and technology. | ○ | ○ | ○ | ○ | ○ | ○ |

* 15. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| This activity is meaningful | ○ | ○ | ○ | ○ | ○ | ○ |
| By participating in this project, I meet people with similar interests. | ○ | ○ | ○ | ○ | ○ | ○ |
| this project is an opportunity to socialize with other people | ○ | ○ | ○ | ○ | ○ | ○ |
| By participating in this project I get to spend more time in nature | ○ | ○ | ○ | ○ | ○ | ○ |
| Participating in this project provided me a teaching opportunity | ○ | ○ | ○ | ○ | ○ | ○ |

* 16. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| I can compete with others in this project | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| This activity taught me new skills or knowledge | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| I can satisfy my need for fame by doing this activity | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By contributing to this project I can make scientific knowledge accessible to the public | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By doing this activity I can help others. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 17. How have your expectations about participating in the project been fulfilled?

| | don't agree at all | rather don't agree | undecided | rather agree | fully agree | don't know/not applicable |
|---|---|---|---|---|---|---|
| By contributing to this project I get to have some good times with my family. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By participating in this project I get to spend more time outdoors | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| By contributing to this project I can contribute to conservation of the environment | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| This activity is fun for me | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

Citizen Science Motivation EN

Thank you for participating to this survey.

18. What is you gender?

- ◯ male
- ◯ female
- ◯ other/prefer not to answer

19. What is your age?

- ◯ Below 18
- ◯ 18-20
- ◯ 21-29
- ◯ 30-39
- ◯ 40-49
- ◯ 50-59
- ◯ Above 60

20. What is the highest degree of education you have completed?

- ◯ Less than primary school
- ◯ Primary school
- ◯ Secondary school level
- ◯ High school / Matura
- ◯ Bachelor
- ◯ Master/Diploma
- ◯ Promotion/PhD/Doctorate

other (please state)

21. What is your country of residence?

22. Do you have comments about the questionnaire and/or the project?

23. Are you interested in the "Hydrosommelier - Bottle" or the result of this study?

- [ ] Please add my e-mail adress to the list of potential recipients of a "Hydrosommelier-Bottle".
- [ ] I am interested in receiving information about the outcomes of this study.

E-mail adress:

For feedback and questions, you can contact me directly: simon.etter@geo.uzh.ch
For more information about my PhD work: www.crowdwater.ch

Citizen Science Motivation

Danke, dass Sie sich die Zeit nehmen!

**Sie wurden angefragt an dieser Umfrage teilzunehmen, da Sie Teil von mindestens einem der Citizen Science Projekte auf der SPOTTERON-Plattform (www.spotteron.net) sind.**

**Ich werde unter allen Teilnehmern, die die Umfrage abschliessen 10 "Hydrosommelier-Flaschen" verteilen. Um teilzunehmen, haben sie am Ende der Umfrage die Möglichkeit ihre E-Mail Adresse anzugeben.**

**Die Umfrage ist Teil meiner Doktorarbeit an der Universität Zürich. Ich arbeite am Projekt CrowdWater und erforsche unter anderem die Motivation von Citizen Scientists. In der nachfolgenden Umfrage stelle ich Ihnen einige Fragen in Form von Statements, die Sie auf einer Skala von 1 (stimme überhaupt nicht zu) bis 5 (stimme vollständig zu) bewerten müssen, je nachdem wie gut oder schlecht diese auf Sie zutreffen.**

**Im ersten Teil (Fragen 5-11) werden Sie nach den Gründen gefragt, weshalb Sie sich zur Teilnahme bei einem oder mehreren Citizen Science Projekten auf der SPOTTERON-Plattform entschieden haben. Falls Sie an mehreren Projekten teilnehmen, wählen Sie bitte dasjenige, zu welchem Sie als erstes beigetragen haben. Bitte berücksichtigen Sie im ersten Teil nicht, ob Ihre Erwartungen erfüllt wurden.**

**Im zweiten Teil (Fragen 12-17) werden Sie dann gefragt, wie gut diese Erwartungen erfüllt wurden, die sie möglicherweise an das Projekt hatten. Sie können auch angeben wie fest sie einem Punkt zustimmen, wenn Sie diese Erwartung anfangs nicht hatten. Berücksichtigen Sie dafür, was Sie heute empfinden.**

**Sie würden mir sehr helfen, wenn Sie am Ende vier kurze Fragen zu Ihrer Person beantworten würden.**

**Ihre Antworten werden anonym erfasst und abgespeichert. Die Umfrage dauert ca. 10 Minuten.**

**Vielen Herzlichen Dank für Ihre Hilfe!**
**Simon Etter**
**(simon.etter@geo.uzh.ch)**

* 1. Um eine Verknüpfung dieses Fragebogens mit einer Folgestudie zu ermöglichen, bitten wir Sie einen persönlichen Code zu erstellen. Dieser besteht aus uns unbekannten Parametern die keine Rückschlüsse auf ihre Person zulassen: Die ersten zwei Buchstaben des Vornamens ihrer Mutter und ihres Vaters und die zweistellige Nummer des Tages von ihrem Geburtsdatum.

Beispiel:

Vorname Mutter: **An**ita

Vorname Vater: **Wa**lter

Geburtsdatum: **01**.02.1980

Beispiel für den persönlichen Code: **AnWa01**

* 2. In welchem SPOTTERON Projekt nehmen Sie Teil (mehrere Antworten möglich)?

- ☐ CrowdWater
- ☐ Naturkalender ZAMG
- ☐ Naturkalender NÖ
- ☐ Naturkalender Steiermark
- ☐ Roadkill
- ☐ Forschen im Almtal
- ☐ Global 2000 DRECKSPOTZ
- ☐ StreetArt
- ☐ Waldrapp
- ☐ Was geht ab?
- ☐ Fågelbär
- ☐ GEFABE
- ☐ Fjällkalendern
- ☐ anderes Projekt

3. Wann sind Sie dem Projekt beigetreten?

- ○ vor ein paar Tagen
- ○ vor Wochen
- ○ vor Monaten
- ○ in der zweiten Hälfte vom Jahr 2017
- ○ in der erste Hälfte vom Jahr 2017
- ○ im Jahr 2016
- ○ im Jahr 2015
- ○ vor 2015

4. Wie oft tragen sie zum Projekt bei?

- ○ mehr als einmal täglich
- ○ einmal täglich
- ○ alle paar Tage
- ○ wöchentlich
- ○ alle paar Wochen
- ○ monatlich
- ○ weniger als monatlich
- ○ Ich habe nur wenige Male beigetragen (1-3 Mal)
- ○ Ich habe noch nie beigetragen

Citizen Science Motivation

## Wieso haben Sie sich zur Teilnahme an einem Citizen Science Projekt entschieden?

* 5. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich will neue Fähigkeiten oder Wissen erlernen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch Freiwilligenarbeit fühle ich mich wichtig. | ○ | ○ | ○ | ○ | ○ | ○ |
| Andere Leute, die ich kenne, machen mit. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich will eine schöne Zeit mit der Familie/Freunden verbringen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Andere Leute denken positiv über mein Beitragen zu diesem Projekt. | ○ | ○ | ○ | ○ | ○ | ○ |

* 6. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich möchte zur Zukunft der Menschheit beitragen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich will Zeit in der Natur verbringen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich möchte wissenschaftliches Wissen der Allgemeinheit verfügbar machen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich strebe nach Ruhm. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Mich interessiert das Thema dieses Projekts. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 7. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich interessiere mich für Wissenschaft und Technik. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Jemand hat von mir verlangt an diesem Projekt teilzunehmen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Diese Aktivität ist mit einem Hobby verwandt, das ich bereits habe. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich möchte die Welt zu einem besseren Ort machen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Es ist eine Gelegenheit andern etwas beizubringen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

* 8. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich will Spass haben. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich will Teil dieser Gemeinschaft von Freiwilligen sein. | ○ | ○ | ○ | ○ | ○ | ○ |
| Mein Glaube und/oder meine Werte haben mich zur Teilnahme motiviert. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich möchte Anerkennung erhalten. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich will etwas Bedeutsames machen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 9. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich möchte zur Wissenschaft beitragen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich mache diese Aktivität, weil ich gerne helfe. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich möchte physisch aktiv sein. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich möchte unter die Leute kommen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich möchte mein Wissen und meine Erfahrung teilen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 10. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich möchte mehr Zeit draussen verbringen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich möchte zum Umweltschutz beitragen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich strebe danach mich selbst herauszufordern. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich will meinen Ruf verbessern. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich erwarte etwas als Gegenleistung. | ○ | ○ | ○ | ○ | ○ | ○ |

* 11. Was war der Grund für Ihre Teilnahme am Projekt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich möchte Leute mit ähnlichen Interessen treffen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich möchte das öffentliche Bewusstsein für dieses Thema erhöhen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich möchte meinen sozialen Status verbessern. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich messe mich gerne mit andern. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| Ich möchte zum Wissen über dieses Thema beitragen. | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

Citizen Science Motivation

## Wie wurden ihre Erwartungen durch die Teilnahme am Projekt erfüllt?

* 12. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Durch die Teilnahme am Projekt kann ich das öffentliche Bewusstsein über dieses Thema verbessern. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt kann ich zum Wissen über dieses Thema beitragen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt kann ich mein Wissen und meine Erfahrungen teilen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt bin ich physisch aktiv. | ○ | ○ | ○ | ○ | ○ | ○ |
| Diese Aktivität hilft mir zur Wissenschaft beizutragen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 13. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Diese Aktivität hilft mir meinen Ruf zu verbessern. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch das Beitragen zum Projekt kann ich etwas für die Zukunft der Menschheit tun. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich erhalte Anerkennung für meine Beiträge zum Projekt. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die freiwillige Arbeit in diesem Projekt fühle ich mich wichtig. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich geniesse diese Aktivität. | ○ | ○ | ○ | ○ | ○ | ○ |

* 14. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Durch das Beitragen zu diesem Projekt kann ich die Welt zu einem besseren Ort machen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch das Helfen in diesem Projekt handle ich entsprechend meines Glaubens und/oder meiner Werte. | ○ | ○ | ○ | ○ | ○ | ○ |
| Diese Aktivität fordert mich heraus. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch diese Aktivität kann ich meinen sozialen Status verbessern. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch mein Beitragen zum Projekt, kann ich mein Interesse in Wissenschaft und Technik befriedigen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 15. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Diese Aktivität ist bedeutsam. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme am Projekt treffe ich Leute mit ähnlichen Interessen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Dieses Projekt gibt mir die Möglichkeit unter die Leute zu kommen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt kann ich mehr Zeit in der Natur verbringen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Die Teilnahme an diesem Projekt gibt mir die Möglichkeit andern etwas beizubringen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 16. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Ich kann mich im Projekt mit anderen messen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch diese Aktivität habe ich neue Fähigkeiten oder neues Wissen erlangt. | ○ | ○ | ○ | ○ | ○ | ○ |
| Ich kann mein Streben nach Ruhm durch meine Teilnahme in diesem Projekt befriedigen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt, kann ich wissenschaftliches Wissen der Öffentlichkeit zugänglich machen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch diese Aktivität kann ich andern helfen. | ○ | ○ | ○ | ○ | ○ | ○ |

* 17. Wurden die folgenden Punkte durch die Teilnahme am Projekt erfüllt?

| | stimme überhaupt nicht zu | stimme eher nicht zu | unentschieden | stimme eher zu | stimme vollständig zu | Weiss nicht/keine Angabe |
|---|---|---|---|---|---|---|
| Durch das Beitragen zum Projekt kann ich eine schöne Zeit mit der Familie/mit Freunden verbringen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt, kann ich mehr Zeit draussen verbringen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Durch die Teilnahme an diesem Projekt kann ich zum Umweltschutz beitragen. | ○ | ○ | ○ | ○ | ○ | ○ |
| Diese Aktivität macht mir Spass. | ○ | ○ | ○ | ○ | ○ | ○ |

## Vielen Dank für Ihre Teilnahme!

18. Was ist ihr Geschlecht?

- ○ Männlich
- ○ Weiblich
- ○ andere/keine Angabe

19. Wie alt sind Sie?

- ○ Unter 18
- ○ 18-20
- ○ 21-29
- ○ 30-39
- ○ 40-49
- ○ 50-59
- ○ Über 60

20. Was ist der höchste Bildungsgrad, den Sie bisher erlangt haben?

- ○ Weniger als Grundschule/Primarschule/Volksschule
- ○ Primarschule/Grundschule
- ○ Sekundarschulabschluss
- ○ Matura bzw. Abitur
- ○ Bachelor
- ○ Master/Diplom
- ○ Promotion/PhD/Doktorat

Sonstiges (bitte angeben)

21. In welchem Land leben Sie?

22. Haben Sie Kommentare bezüglich dieses Fragebogens oder des Projekts?

23. Sind sie interessiert an einer "Hydrosommelier"-Flasche oder an den Resultaten dieser Studie?

- [ ] Bitte fügen Sie meine E-Mail Adresse zur Liste der potentiellen Empfänger einer "Hydrosommelier"-Flasche hinzu.
- [ ] Ich möchte über die Resultate dieser Studie informiert werden.

E-Mail Adresse:

Für Feedback und Fragen können Sie sich direkt an mich wenden: simon.etter@geo.uzh.ch
Für weitere Infos bezüglich meiner Doktorarbeit: www.crowdwater.ch

# Paper III

HYDROLOGICAL SCIENCES JOURNAL
https://doi.org/10.1080/02626667.2019.1578966

SPECIAL ISSUE: HYDROLOGICAL DATA: OPPORTUNITIES AND BARRIERS

# Accuracy of crowdsourced streamflow and stream level class estimates

Barbara Strobl [a], Simon Etter [a], Ilja van Meerveld [a] and Jan Seibert [a,b]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

**ABSTRACT**
Streamflow data are important for river management and the calibration of hydrological models. However, such data are only available for gauged catchments. Citizen science offers an alternative data source, and can be used to estimate streamflow at ungauged sites. We evaluated the accuracy of crowdsourced streamflow estimates for 10 streams in Switzerland by asking citizens to estimate streamflow either directly, or based on the estimated width, depth and velocity of the stream. Additionally, we asked them to estimate the stream level class by comparing the current stream level with a picture that included a virtual staff gauge. To compare the different estimates, the stream level class estimates were converted into streamflow. The results indicate that stream level classes were estimated more accurately than streamflow, and more accurately represented high and low flow conditions. Based on this result, we suggest that citizen science projects focus on stream level class estimates instead of streamflow estimates.

**ARTICLE HISTORY**
Received 19 June 2018
Accepted 22 November 2018

**EDITOR**
A. Castellarin

**GUEST EDITOR**
C. Cudennec

**KEYWORDS**
citizen science; crowdsourcing; stream level; stream level class; streamflow; accuracy; CrowdWater

## 1 Introduction

Streamflow data are important for many aspects of river management, including water allocation and the reduction of flood hazards. Streamflow data are also important for the calibration of hydrological models to predict floods and droughts or the impacts of climate change. Most hydrological models need at least a certain amount of data to be properly "tuned" to a particular catchment (Beven 2012).

Three important aspects define the usability of streamflow data: accuracy, spatial coverage and temporal resolution. Conventional streamflow gauging stations can provide detailed information with high accuracy and temporal resolution, but the spatial coverage is limited. While data from gauging stations are considered accurate, the data can still contain substantial errors due to sensor errors, interpolation and extrapolation of the rating curve and cross-section instability (McMillan *et al.* 2012). Typical relative errors for streamflow are ±50–100% for low flows and ±10–20% for medium or high flows (still within the streambank) (McMillan *et al.* 2012). Similar values were derived by Westerberg *et al.* (2011), who mentioned rating curve related errors of −60% to +90% for low flows and ±20% for medium to high flows.

The temporal resolution of gauging stations is often high. However, due to financial and logistic constraints, only a few sites have a gauging station, hence the spatial coverage is limited. Furthermore, these stations may not be installed at representative locations or might miss certain types of catchments, especially small headwater streams (Kirchner 2006, Bishop *et al.* 2008). Also relatively few measurement stations are located in developing countries. Thus, for many catchments there are no streamflow data available for water management decisions or model calibration.

Although new wireless sensor network technology provides the possibility to expand the measurement networks, the reality is that, due to budget cuts, observation networks often shrink rather than expand (Kundzewicz 1997, Ruhi *et al.* 2018). For example, Ruhi *et al.* (2018) showed that between 1947 and 2016 the number of streamgauges in river basins in the USA decreased by 21%.

Several studies have focused on the minimum number of measurements required to properly calibrate a hydrological model (Perrin *et al.* 2007, Juston *et al.* 2009, Seibert and Beven 2009, Seibert and McDonnell 2015, Vis *et al.* 2015) and have shown that even a few streamflow measurements can vastly improve the performance of a model (Pool *et al.* 2017). While employees of agencies responsible for national or regional gauging station networks could perhaps take a limited number of additional measurements at a few ungauged streams, it is impossible for them to take measurements

**CONTACT** Barbara Strobl barbara.strobl@geo.uzh.ch
Supplemental data for this article can be accessed here.

at all ungauged streams. An interesting alternative to obtaining streamflow data for more streams is to ask citizen scientists or citizen observers to collect streamflow data.

Citizen science has been used in numerous environmental studies to obtain data with a much higher spatial resolution than is otherwise possible (Dickinson *et al.* 2010, Tulloch *et al.* 2013, Aceves-Bueno *et al.* 2017, Hadj-Hammou *et al.* 2017) and has been used to obtain hydrological data as well (Buytaert *et al.* 2014). For example, citizen science data have been used to fill in spatial and temporal gaps in water quality and stream level data series (Lowry and Fienen 2013, Hadj-Hammou *et al.* 2017) and to obtain groundwater level data across large areas (Little *et al.* 2016). Citizen science could therefore be a complementary approach to collect the stream level and streamflow data that are needed for hydrological model calibration, particularly for the many streams that are currently ungauged. In order to involve as many citizens in data collection as possible and to obtain data for remote areas, approaches are needed to collect these data with very little time and effort and without special equipment.

Despite their potential to complement existing data sources, citizen science data are not without challenges; in particular, the accuracy of crowdsourced data is often discussed (Engel and Voshell 2002, Haklay 2010, See *et al.* 2013, Aceves-Bueno *et al.* 2017). Several studies have examined the accuracy of crowdsourced hydrological data (Turner and Richter 2011, Rinderer *et al.* 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer *et al.* 2015, Le Coz *et al.* 2016, Little *et al.* 2016, Weeser *et al.* 2018). Lowry and Fienen (2013) found promising results in terms of the accuracy of stream level data from participants who read the level from a staff gauge in a stream close to a hiking path. The root mean square error (RMSE) of the crowdsourced stream level data was approximately 5 mm, which was almost as good as that of pressure transducer data. They concluded that the level of accuracy "*is encouraging since no training was given to the citizen scientists*" (Lowry and Fienen 2013, p. 155). In a similar study by Weeser *et al.* (2018) in Kenya, data collected by citizens were comparable to those of conventional data loggers, although they had a low temporal resolution. Little *et al.* (2016) provided volunteers with equipment to measure the water level in their own wells. They found that the absolute difference of the well readings ranged from 2 to 11 mm and concluded that "*community-based groundwater monitoring provides an effective and affordable tool for sustainable water resources management*" (Little *et al.* 2016, p. 317). Peckenham and Peckenham (2014) analysed groundwater quality data collected by students and concluded that the accuracy varied, but "*it is possible to make precise and accurate measurements consistent with the methods specifications*" (Peckenham and Peckenham 2014, p. 1477).

However, these previous hydrological citizen science studies are not easily scalable to many sites because they require the installation of staff gauges or other instrumentation. Therefore, it is useful to also develop and test citizen science approaches to collect streamflow or stream level data that do not require equipment or the installation of staff gauges, but these new citizen science tasks should be designed "*with the skill of the citizens in mind*" (Aceves-Bueno *et al.* 2017, p. 287). It is likely that many citizens who frequently pass by streams notice high and low flows throughout the seasons. These frequently visited locations could be turned into locations for streamflow or stream level class observations if citizens can accurately estimate streamflow or stream level classes.

Testing the accuracy of citizen science data before starting a citizen science project is crucial for every citizen science project. This ensures that the data collected are sufficiently accurate for the purpose of the project and avoids unnecessarily burdening citizens with tasks that result in data that are in hindsight of limited value due to data accuracy issues. The objective of this study was, therefore, to determine what types of parameters related to streamflow citizens can estimate accurately. We asked 517 citizens to estimate both the streamflow and stream level class and assessed whether one can be estimated more accurately than the other by calculating the corresponding streamflow for each stream level class estimate. Accuracy is defined here as the difference between the estimated value and the measured value, as well as the frequency of extreme outliers. The specific research questions for this study were:

(1) How well can stream level class, streamflow and the different factors of streamflow (width, depth, flow velocity) be estimated by citizens?
(2) To what extent do stream size and flow conditions affect the accuracy of the crowdsourced data?

## 2 Methodology

### 2.1 Basic approach and study sites

We conducted 16 field surveys where we asked people to estimate the streamflow, as well as the average width, depth and velocity of the stream, and the stream level class. For the surveys, we selected 10 locations (Table 1; see also Supplementary material, Fig. S1) where we

**Table 1.** Information on the streams where the field surveys took place. Size classes XS: ≤1 $m^3$/s; S: >1–50 $m^3$/s, M: >50–200 $m^3$/s and L: >200 $m^3$/s. A map with the survey locations is given in the Supplementary material (Fig. S1). Survey dates given as dd.mm.yyyy.

| Stream | Size | | Date of survey | No. of participants, *n* | Streamflow ($m^3$/s) | Source for measured streamflow* | Approx. distance to virtual staff gauge (m) | Comments |
|---|---|---|---|---|---|---|---|---|
| Chriesbach (Zurich) | XS | | 29.09.2017 | 30 | 0.38 | Salt dilution | 5 | BSc students: no direct streamflow estimates |
| Hornbach (Zurich) | XS | | 19.02.2017 | 33 | 0.134 | Salt dilution | 8 | |
| Irchel (Zurich) | XS | | 11.03.2017 | 25 | 0.01 | Salt dilution | 1 | |
| Glatt (Zurich) | S | | 29.09.2017 | 31 | 2.8 | WWEA, station: 533 | 11 | BSc students: no direct streamflow estimates |
| Magliasina (Magliaso) | S | | 28.04.2017 | 40 | 16 | FOEN, station: 2461 | 14 | High-school students: no stream level class estimates |
| Schanzen-graben (Zurich) | S | | 01.04.2017 | 31 | 2.6 | Salt dilution | 16 | |
| Sihl (Zurich) | S | 1 | 18.02.2017 | 33 | 7 | FOEN, station: 2176 | 32 | Low flow |
| | | 2 | 26.07.2017 | 31 | 28 | | | High flow |
| Töss (Winterthur) | S | | 12.03.2017 | 35 | 9 | WWEA, stations: 518, 520 and 581 | 29 | Interpolation between three nearby stations for reference value |
| Limmat (Zurich) | M | 1 | 29.10.2016 | 38 | 59 | FOEN, station: 2099 | 7 | No stream level class estimates |
| | | 2 | 08.04.2017 | 27 | 83 | | | |
| | | 3 | 02.06.2017 | 31 | 107 | | | |
| | | 4 | 09.07.2017 | 44 | 75 | | | PhD students Low flow |
| | | 5 | 13.11.2017 | 31 | 222 | | | High flow |
| Aare (Brugg) | L | 1 | 07.01.2017 | 27 | 108 | FOEN, station: 2016 | 53 | Low flow |
| | | 2 | 10.05.2017 | 30 | 389 | | | High flow |

* The measured streamflow data were obtained from the Federal Office of the Environment (FOEN; http://hydrodaten.admin.ch/), the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA; www.hydrometrie.zh.ch/) or by salt dilution gauging (Salt dilution).

expected enough people to pass by and have time for the survey. We divided the streams into four different size classes (XS, S, M, L) based on the mean annual streamflow, and, when long-term time series were not available, based on the available measurements:

- XS (Chriesbach, Hornbach and Irchel): ≤1 $m^3$/s,
- S (Glatt, Magliasina, Schanzengraben, Sihl and Töss): >1–50 $m^3$/s,
- M (Limmat): >50–200 $m^3$/s, and
- L (Aare): >200 $m^3$/s.

To analyse whether the flow conditions affect the accuracy of the estimates, surveys were conducted under high and low flow conditions for three streams: Aare (L), Limmat (M) and Sihl (S).

The aim of the surveys was to get a sufficient number of streamflow estimates for a specific stream on a specific day (our aim was 30 participants per survey to assure statistical significance; Field *et al.* 2013). We therefore used a logistically simple sampling strategy, whereby we personally approached passers-by (similar to Breuer *et al.* 2015) and asked if they would complete the 5-minute survey (i.e., we did not use a targeted approach to capture responses of a representative group of citizens). No data were collected on the percentage of passers-by who participated, but we estimate that about every third person we approached agreed to participate in our survey. In addition, we asked high-school (Magliasina) and university students (Chriesbach, Glatt and Limmat) to fill out the survey during excursions. All surveys took place between October 2016 and September 2017. In total, we received 517 complete surveys: 372 passers-by, 61 participants from a university geography bachelor student excursion (Glatt and Chriesbach), 40 from a high-school student excursion (Magliasina) and 44 from a summer school for PhD students from fields ranging from physics to social sciences (Limmat) (see Table 1). During the group excursions we emphasized the need for individual estimates and limited discussions between the students for the duration of the survey.

The age distribution of all 517 participants corresponds to that of the inhabitants of Zurich (where most field surveys were conducted), although there were fewer participants over the age of 60 (13% of the participants *vs* 19% of the population in Zurich; see Supplementary material, Fig. S2(c) and (d)) (Statistik Stadt Zürich 2017). Also a large number of participants were university educated, roughly 48% compared to 16% of the population in Zurich (Fig. S2(b)) (Statistik Stadt Zürich 2017). There was an almost equal split between male and female participants (Fig. S2(a)).

### 2.2 Streamflow estimation

Participants were first asked to estimate the streamflow directly. For this direct estimate, we asked them to estimate the flow in $m^3/s$, or in L/s for the very small streams (XS). This directly estimated streamflow value is referred to as $Q_{direct}$. This task, understandably, proved to be difficult for some participants because streamflow quantification was difficult and they were unfamiliar with the units. A few participants refused to answer this question, even with a bit of prompting. Some decided to guess, even though they thought it was unlikely to be a realistic value and others deduced on their own that they could estimate the width, mean depth and flow velocity to get an approximate value.

After this initial guess of the streamflow, we explained to the participants that it is possible to estimate the individual factors (width, mean depth and flow velocity) and to derive the streamflow by multiplying these values (Equation (1)). The participants were then asked to estimate the average width, mean depth and velocity of the stream. We also asked them to classify the streambed material. Equation (1) was used to calculate the streamflow using these factors:

$$Q_{factor} = w \cdot d \cdot v \cdot k \tag{1}$$

where $Q_{factor}$ is the estimated streamflow ($m^3/s$), $w$ is the estimated width (m), $d$ is the estimated mean depth (m), $v$ is the estimated surface flow velocity (m/s) and $k$ is the correction factor to obtain the average velocity from the surface velocity. While some participants still found the quantification difficult, they were more familiar with these units, compared to $m^3/s$ or L/s. Often a value of 0.85 is used for the correction factor $k$ (Welber *et al.* 2016); but it can also be estimated using the logarithmic velocity distribution (Prandtl-von Kármán equation) for turbulent flow based on the surface flow velocity, grain size and stream depth (Dingman 2015). This calculated factor for the mean flow velocity varied for the different estimates of the participants (even for the same stream). For two-thirds of all estimates, the calculated velocity factor was not within the typical range of 0.71–0.95 (Welber *et al.* 2016) due to an unrealistic ratio between the estimated average water depth and estimated streambed roughness. Values lower than 0.71 were adjusted to 0.71 (52% of estimates) and values over 0.95 were adjusted to 0.95 (1% of estimates). When no estimate for streambed roughness was available (this happened only occasionally, except for the entire field survey at Magliasina), the typical velocity correction factor of 0.85 was used (including the participants at Magliasina this corresponds to 13% of all estimates). During the university excursion at the Glatt and Chriesbach, we did not ask for direct stream estimates because most geography bachelor students would likely have applied the indirect estimation method ($Q_{factor}$) because of lectures on streamflow during their education.

To assess the accuracy of crowdsourced streamflow data, the streamflow estimates were compared to measured streamflow data. Streamflow was measured before or after the surveys (Chriesbach, Hornbach, Irchel and Schanzengraben) or obtained from official gauging station data when these were located near the survey location (Aare, Limmat, Magliasina and Sihl, stations of the Swiss Federal Office for the Environment (FOEN); Glatt and Töss, stations of the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA)) (see Table 1). The methods for the reference measurements for width, mean depth and flow velocity depended on the size and accessibility of the river. These measurements included direct measurements for width and depth with measurement tapes, data on the stream cross-section from FOEN for width and depth (when available), an estimate of the width of the river from Google Maps for wide rivers (Aare and Limmat) and the stick method for flow velocity. Even though these measurements are likely also affected by errors, they were assumed to be the "true" data to which the citizen science estimates could be compared. We assumed that the uncertainty for the measured values is 10% for streamflow (Pelletier 1988), 0.5% for width and 1–3% for depth (Herschy 1971) and roughly 10% for flow velocity (based on our own measurements).

### 2.3 Stream level class estimation

We also asked participants to estimate the stream level class. Stream level refers to the height of the water in a stream. A stream level class means that this height is expressed on a discrete scale of classes, rather than on a continuous scale. Stream level class data only provide information about whether the stream level is higher or lower than previously, but earlier studies have shown that stream level class data are useful for hydrological model calibration (van Meerveld *et al.* 2017). Thus, the participants were not asked to estimate the stream level in centimetres but to estimate the stream level class. The participants compared the current stream level with a photo of the same stream (taken at an earlier time) with a digitally inserted staff gauge with 10 level classes (Fig. 1, also Supplementary material, Section S2). The staff gauge was scaled so

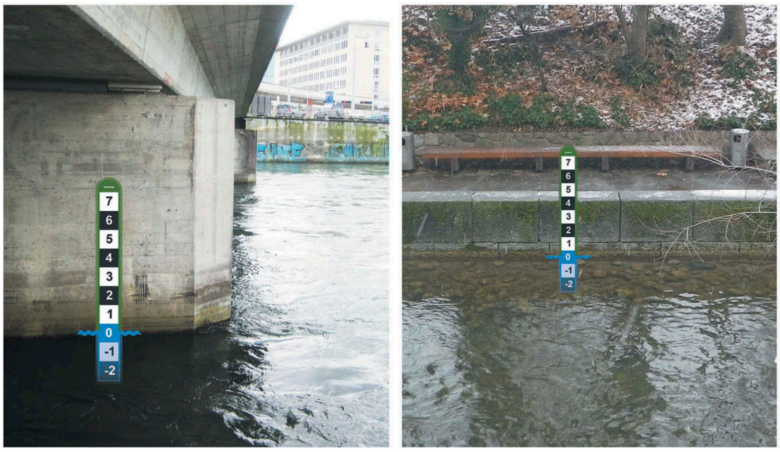



Figure 1. Example of a virtual staff gauge in the pictures used for the surveys at Limmat (left) and Schanzengraben (right). Photographs taken on 29.06.2016 when the streamflow was 165 $m^3/s$ (Limmat) and on 05.01.2017 (unknown streamflow; Schanzengraben). For the dates and the flow conditions during the surveys see Table 1.

that the highest class represented the highest in bank flood level and the lowest class represented the likely lowest stream level. The height of the classes is arbitrary and varied for each location, depending on the size of the river and how the virtual staff gauge was placed in the picture. A small staff gauge would have a higher resolution, but the stream level for very high and low flows may be above or below the staff gauge, whereas a large staff gauge would imply a lower resolution of the observations as the stream level would fluctuate across fewer classes. In this study we tried to place the staff gauges so that the staff gauge covered both high and low in bank flows. The number of classes was a compromise between resolution and usability. A larger number of classes provides higher resolution data but also makes it more difficult (or even impossible) for participants to determine the stream level class. Based on a previous model, study model calibration results do not improve much when more than five stream level classes are used (van Meerveld *et al.* 2017). The number of 10 classes was chosen to ensure observable stream level fluctuations even in cases where the virtual staff gauge is placed so that some classes are never or very rarely reached. The correct stream level class value was determined by us by carefully choosing appropriate references and individually (but unanimously) deciding on the correct stream level class.

For the Limmat, results are given for all five field surveys for streamflow, but stream level class estimates are given for only four surveys because a slightly different virtual staff gauge was used for the first survey.

### 2.4 Data analyses

To be able to compare the accuracy of the streamflow estimates for different streams, relative estimates (in percent) were calculated by dividing the streamflow estimate by the measured value (i.e., considered true value). A value of 100% corresponds to a perfect estimate, smaller values represent an underestimation and larger values represent an overestimation. The quality of the data was then assessed by statistical measures, such as the interquartile range and median. In addition, we determined the number of outliers as they are likely disinformative for model calibration (Beven and Westerberg 2011) and can be worse than having no data. Even though filters can be used to remove outliers in citizen science data, in practice, it may be difficult to filter out all outliers. All relative estimates below 50% and above 150% were considered to be outliers.

For comparison between streamflow and stream level class estimates, stream level classes and the errors in this classification were converted to an equivalent streamflow ($m^3/s$), named $Q_{level}$ in the remainder of the manuscript. For the stream locations with a nearby FOEN gauging station (Sihl, Limmat, Aare), the classes of the virtual staff gauge were converted to a metric value by determining the stream depth that corresponded to each stream level class (i.e., mid-point and upper and lower stream level for each class) and using the FOEN rating curve to convert these stream levels to a streamflow estimate. For the sites where no rating curve was available (Hornbach, Irchel, Schanzengraben and Töss), additional measurements of the stream profile and water surface slope (estimated based on the slope of the streambed) were used to estimate the streamflow for each stream level class using the Manning-Strickler formula (Manning 1891). This curve was fitted to the streamflow measured on the day of the surveys by adjusting the roughness coefficient within predefined boundaries based on the streambed material. The roughness coefficient used for the Manning-Strickler formula introduces some subjectivity and thereby likely increases the uncertainty of the conversion of the stream level class to streamflow compared to FOEN rating curve measurements. Since the stream level classes represent a range of values rather than just one value, the streamflow was not only calculated for the centre value of the level class, but also the class boundaries to obtain the possible range of streamflow values. The estimates from Chriesbach, Glatt and Magliasina were excluded from this analysis (101 of the 517 estimates) because the relevant data were not collected at the time of the surveys.

The differences in the median relative estimates for the different stream size classes were tested for significance using the Kruskal-Wallis test with the *post hoc* procedure based on Dunn (1964). Differences in the median relative streamflow estimates between high and low flow conditions were tested for significance using the Mann-Whitney test. A p-value of 0.05 was used for all statistical tests, unless otherwise indicated.

## 3 Results

### 3.1 Streamflow estimates

Although there was a large spread in the streamflow estimates, the median values were surprisingly close to the measured streamflow (Figs 2 and 3). Across all surveys the median of the direct streamflow estimates ($Q_{direct}$) was closer to the measured value than the estimate based on the factors ($Q_{factor}$) (median relative estimates of 93 and 80%, respectively, when all surveys were analysed together). However, the interquartile range was smaller for the streamflow calculated from the estimated factors (the first and third quartiles were, respectively, 26 and 309% for $Q_{direct}$ and 39 and 172% for $Q_{factor}$; Fig. 3), meaning that the streamflow estimates were closer to the measured value for the estimates based on the factors.

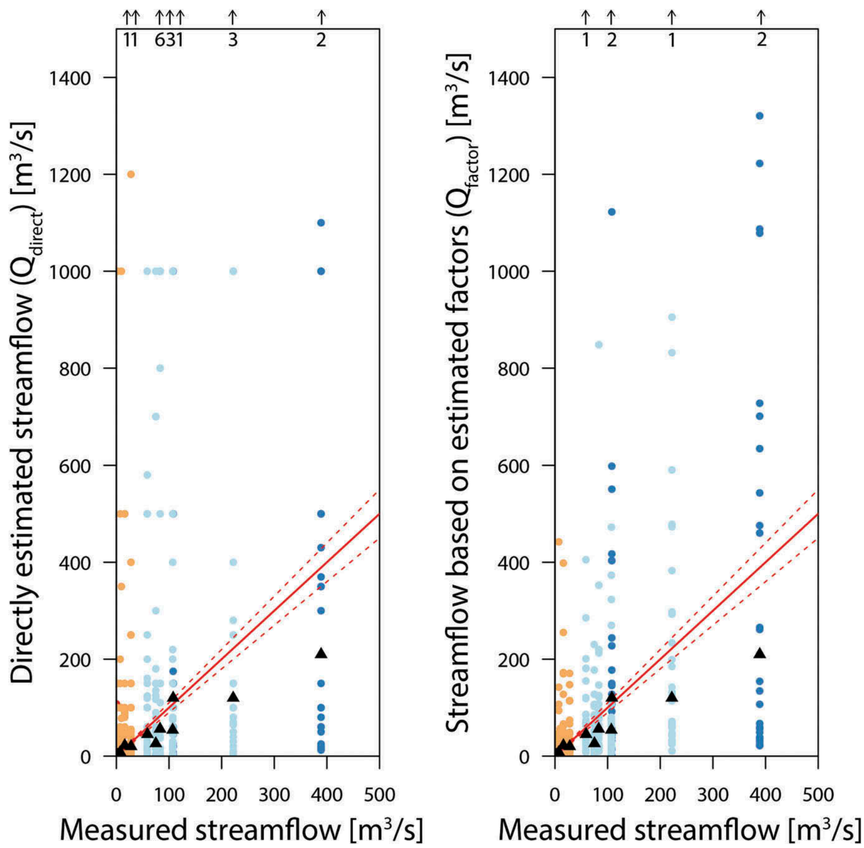



Figure 2. Scatter plots showing the spread of $Q_{direct}$ (left) and $Q_{factor}$ (right) for each field survey. The data points are colour-coded according to the stream size: from left to right, XS to L are red, orange, light blue and dark blue, respectively. ▲: median estimated streamflow per survey; solid and dashed (red) line: the 1:1 line with the 10% uncertainty band. The number at the top of the graph indicates the number of extreme outliers (1–6, not shown).

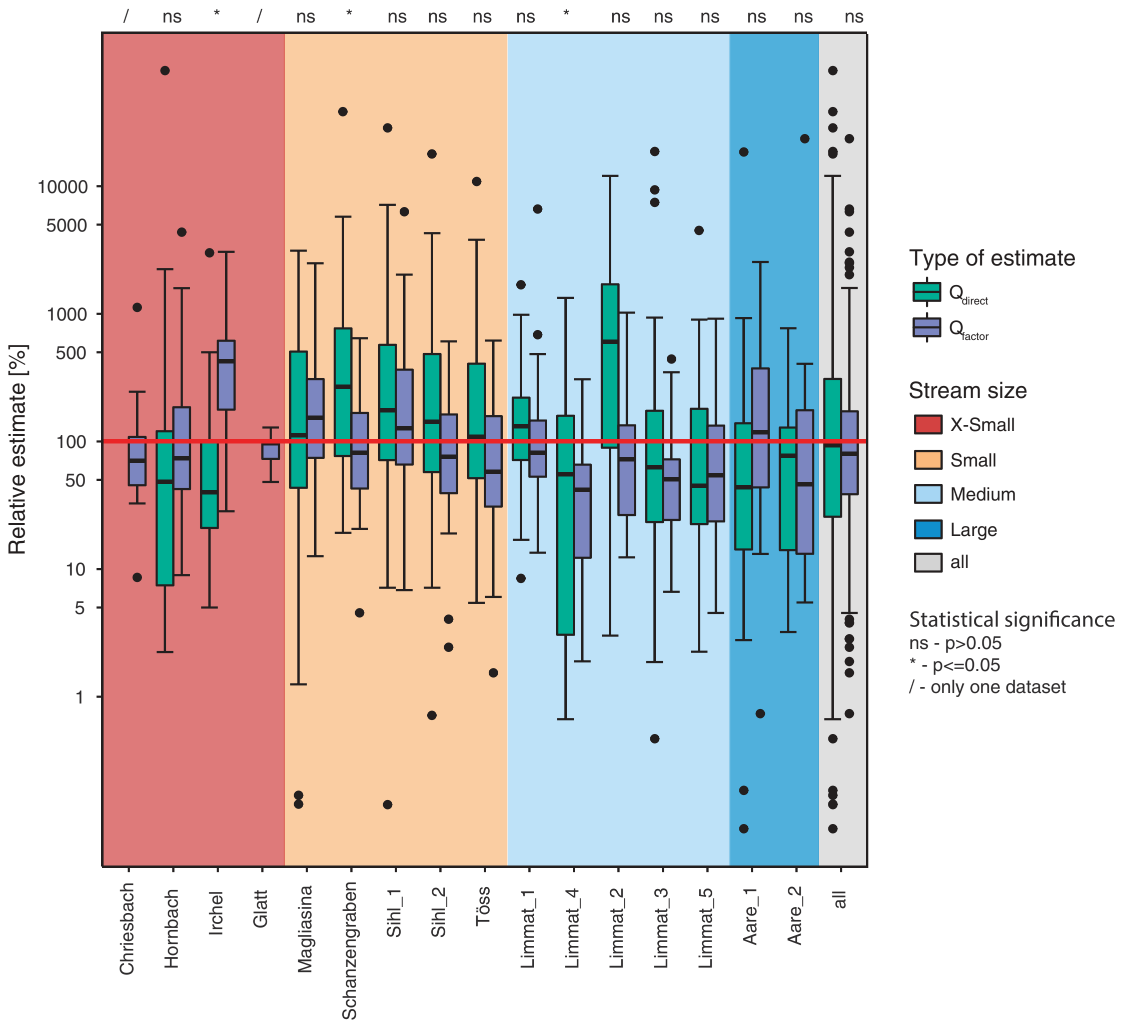



**Figure 3.** Box plots of the relative estimates of streamflow (ratio of estimated *vs* measured streamflow) for $Q_{direct}$ and $Q_{factor}$ for each survey, and for all streams combined (all). Statistical significance, i.e. difference in median relative streamflow estimate for the two methods, is shown across the top. The data for the Sihl, Limmat and Aare are ordered from low to high flow conditions (see Table 1). The box represents the interquartile range, the black line the median, the whiskers extend to 1.5-times the interquartile range below/above the first/third quartile, and the dots represent values beyond 1.5-times the interquartile range. Note the log scale.

The differences between the median estimates of $Q_{direct}$ and $Q_{factor}$ were statistically significant ($p < 0.05$) for three out of the 14 surveys with both $Q_{direct}$ and $Q_{factor}$ estimates, but not for all surveys combined (Fig. 3). Of these three surveys, two had a median estimate for $Q_{direct}$ that was closer to the measured value. The interquartile range was smaller for $Q_{factor}$ for two of the three surveys.

### 3.2 Streamflow factor estimates

There were also numerous outliers for the relative estimates of width, mean depth and flow velocity (Fig. 4). The median relative estimates for the width, depth and flow velocity were all significantly different from each other (Fig. 4). The width was generally underestimated (median relative estimate of 75%, and third quartile of 95% when all stream surveys were analysed together), the mean depth was generally overestimated (median relative estimate of 126% when all stream surveys were analysed together), while the median flow velocity was surprisingly accurate (median relative estimate of 100% when all stream surveys were analysed together). However, the interquartile range suggests that width can be estimated most accurately (interquartile range of relative estimates from 57 to 95% when looking at all surveys together), and mean depth (interquartile range of relative estimates from 86 to 180%) and flow velocity (interquartile range of relative estimates from 57 to 143%) can be estimated less accurately. The percentage of relative estimates below 50% or above 150% shows the

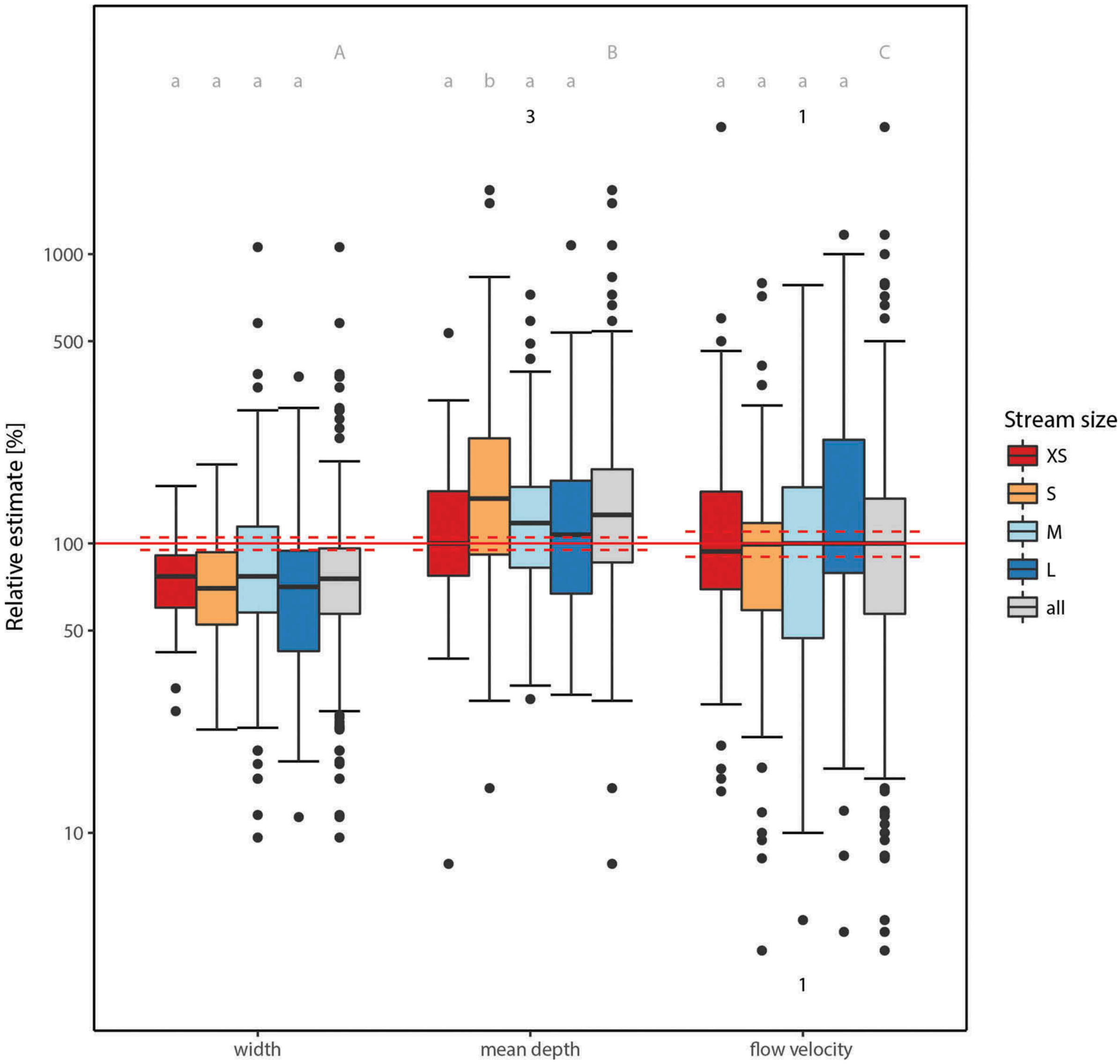



**Figure 4.** Box plots of the relative estimates of width, mean depth and flow velocity for each stream size class and all streams together. Median relative estimates of width, mean depth and flow velocity of all surveys combined were significantly different (indicated by different upper case letters), whereas between stream size classes they were mostly similar (same lower case letters). The solid red line (100%) indicates that the estimate is the same as the measured value; dashed red lines indicate the 5% (width and mean depth) and 10% (flow velocity) uncertainty bands. The numbers above and below the box plots indicate the number of outliers not shown. Note the log scale.

same pattern, with width having fewer outliers (26%) than flow velocity (39%) and mean depth (41%) (Fig. 4). These errors likely occurred due to a misunderstanding of the method.

### *3.3 Stream level class estimates*

About half of the participants (48%) selected the correct stream level class and most of the remaining participants (40%) were off by only one class. There were only a few outliers (13% of participants had an error of two classes or more; the total does not add to 100% due to rounding) (Fig. 5(a)). The largest overestimation was six classes and the largest underestimation was three classes.

### *3.4 Comparison of stream level class and streamflow estimates*

To allow comparison of the streamflow and stream level class estimates, the latter were translated into corresponding streamflow values. These calculated streamflow values had a narrower interquartile range than the streamflow estimates based on the factors (67–157% compared to 30–163% for $Q_{level}$ and $Q_{factor}$,

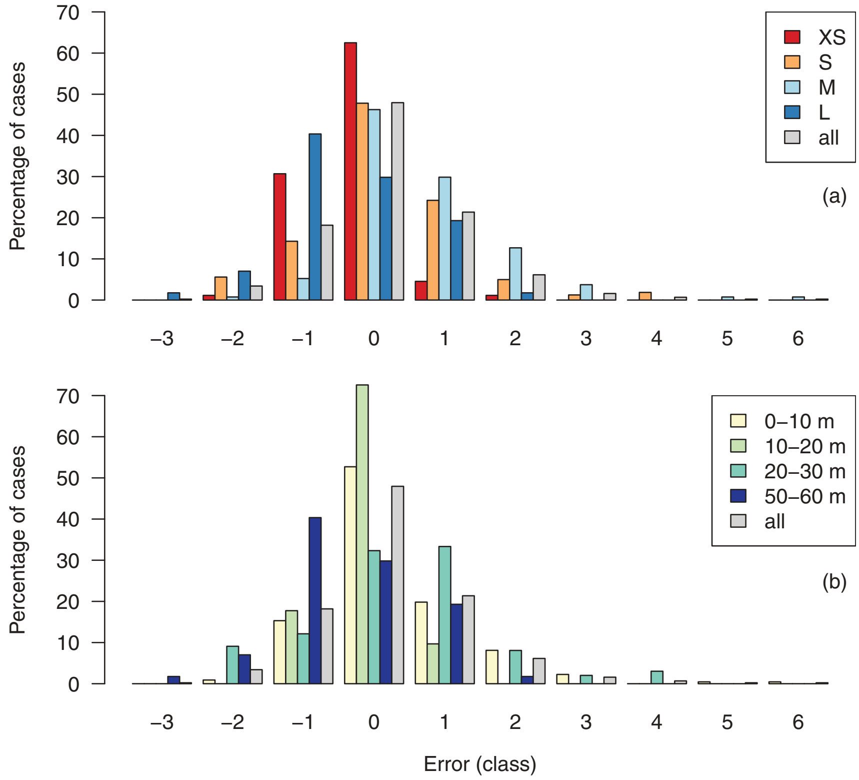



**Figure 5.** (a) Distribution of errors in stream level class estimates (0: no error, −1: one class lower than the actual stream level class, and 1: one class higher than the actual class) for streams of different sizes; and (b) the distance between participant and the virtual staff gauge, as well as all estimates together. There were no surveys where the virtual staff gauge was 30–50 m away from the participants.

respectively, when all estimates are compared together) and also had fewer outliers (see Fig. 6). Only 39% of the streamflow estimates derived from the stream level class estimates (compared to 66% for $Q_{factor}$) were significantly overestimated (relative estimate > 150%) or underestimated (relative estimate < 50%). Furthermore, only 3% of the estimates were more than a factor of 10 "off target" (compared to 11% for $Q_{factor}$). Even when taking the uncertainty in streamflow for the upper and lower stream level class boundaries into account (Fig. 7), the stream level class estimates resulted in streamflow values that were more accurate and had fewer outliers than those determined from the estimated width, mean depth and flow velocity.

Only for the small-sized streams was the interquartile range for streamflow calculated from stream level classes larger than the streamflow determined from the estimated width, depth and flow velocity (Fig. 6). When taking a closer look at the surveys for the different streams, it is clear that mainly the first survey at the Sihl and partly the survey at the Töss caused the large variation in the estimated streamflow from the stream level class data (see Supplementary material, Fig. S3).

## 3.5 Effect of stream size on streamflow and stream level class estimates

### 3.5.1 Streamflow

When estimating streamflow directly ($Q_{direct}$), participants made larger relative errors for the small streams (S; first to third quartile of relative estimates: 55–542%), than for the XS (19–112%), M (23–233%) and L (14–134%) streams. However, general statements on the effect of stream size on the accuracy of streamflow estimates are difficult to make because there were significant differences within each size class as well (Fig. 3).

The interquartile range of the $Q_{factor}$ estimates was significantly smaller for the small (first to third quartile of relative estimates: 49–175%) and medium (27–117%) streams compared to $Q_{direct}$ (Fig. 6). The $Q_{factor}$ estimates were less accurate for XS (interquartile range: 47–293%) and L (17–226%) streams than for S and M streams. For the XS streams this difference is largely based on the estimates from Irchel, where direct streamflow estimates were more accurate than those derived from the estimated factors. For the Hornbach (another XS stream), there was no significant difference between the median relative estimates of $Q_{direct}$ and $Q_{factor}$ (for the Chriesbach

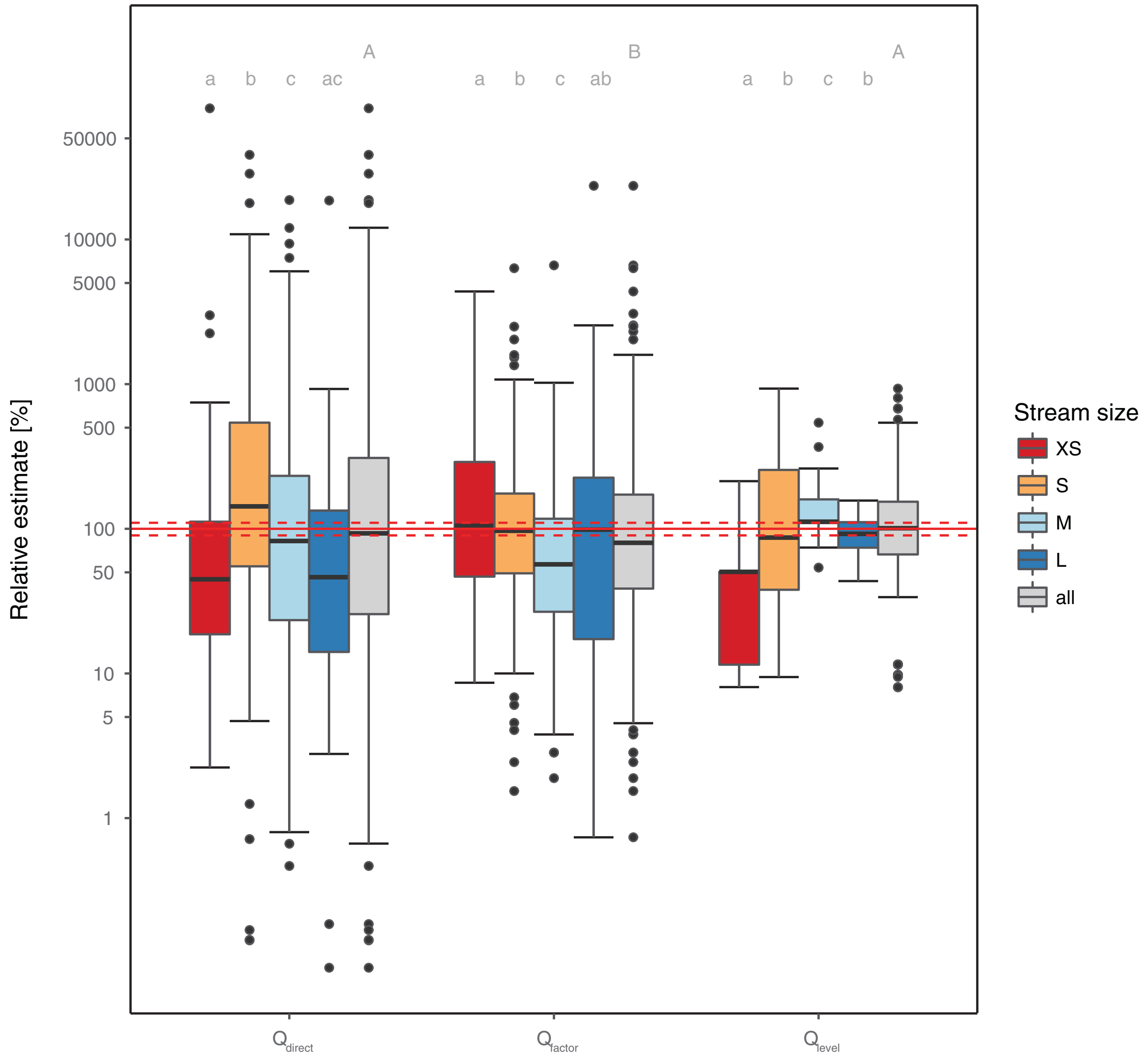



**Figure 6.** Box plot of the relative estimates of $Q_{direct}$, $Q_{factor}$ and $Q_{level}$ for each stream size class and all surveys combined. The statistically significant different medians are indicated by different upper case letters (combined data from all surveys) and different lower case letters (per stream size classes). The solid (red) line at 100% indicates that the estimate is the same as the measured value and the dashed (red) lines indicate the 10% uncertainty band for the measured streamflow.

there was no directly estimated streamflow data). The reasons for this different pattern in the Irchel stream are unknown, but could be due to the lower streamflow in the Irchel stream (0.01 $m^3/s$) compared to the Hornbach (0.13 $m^3/s$).

### 3.5.2 Stream level classes

Stream level class estimates were also analysed according to the distance between the participants and the virtual staff gauge, because the distance was not always related to the stream size. For the Limmat the virtual staff gauge was positioned on a bridge pillar rather than the opposite streambank (Fig. 1).

The stream level class estimates were generally more accurate if the staff gauge was closer to the observer (Fig. 5). For a distance of 0–10 m, 53% of participants selected the correct stream level class, while 35% selected a stream level that was only one class away. For a distance of 10–20 m, no one selected a stream level class more than one class from the true value, and 73% of the participants selected the correct class, while for a distance of 20–30 m, 32% of participants were correct and 45% were one class away. For a distance of 50–60 m, 30% of participants chose the correct stream level class and 60% a neighbouring stream level class (Fig. 5(b)). This is not surprising, as, in cases where the

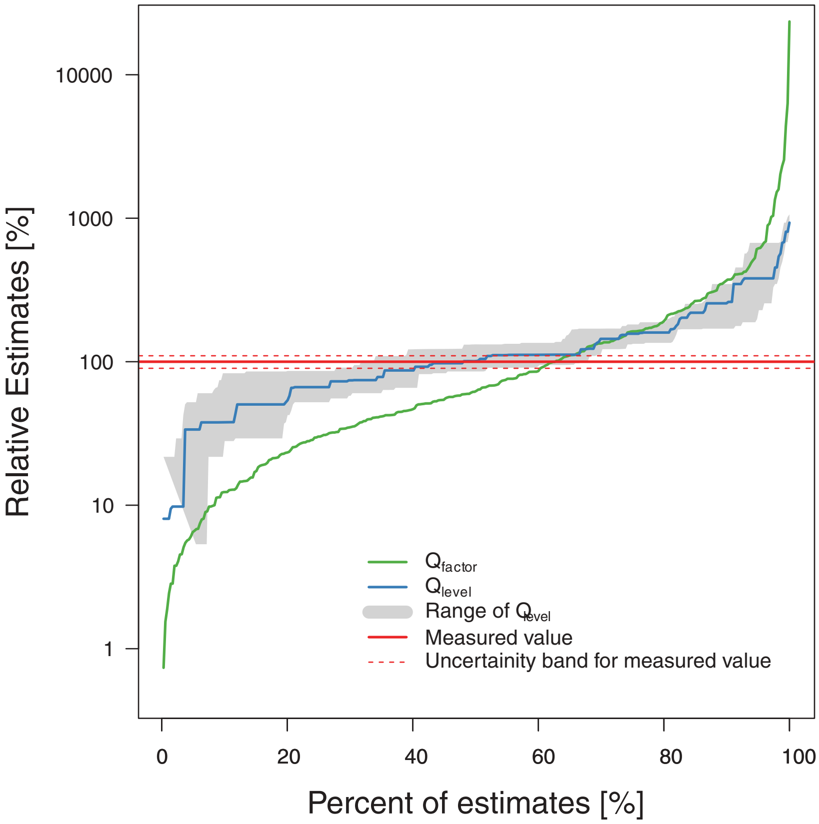



**Figure 7.** Frequency distribution of the relative streamflow estimates for $Q_{factor}$ and $Q_{level}$. The shaded (grey) band indicates the upper and lower streamflow for each stream level class. The lower streamflow for each stream level class does not reach the 0% mark, as there were 18 zero values, which cannot be displayed on a log scale.

virtual staff gauge is far away, it is more difficult to discern the stream level class and the reference, such as stones or other helpful objects, on the streambank.

### 3.6 High *vs* low flow estimates

One issue with hydrological data based on citizen science is the accuracy of the estimated streamflow, but another issue is whether changes in these estimates reflect differences in streamflow over time. Comparison of the estimated streamflow values for the Limmat, Sihl and Aare shows that the median estimated streamflow ($Q_{factor}$) was higher when the flow was higher, but the differences were not sufficient to fully reflect the increased streamflow (Fig. 8) and were not significant for the Aare (Fig. 8(b) and (c)). For the Limmat there were significant differences between the surveys, but these differences did not correspond fully to the measured values, as participants underestimated both high and low flow and the differences of estimates between the surveys were seemingly random regardless of high or low flow (Fig. 8(a)).

The variations in streamflow were better represented by the streamflow derived from the stream level class estimates ($Q_{level}$; Fig. 8(d)–(f)), for which the median estimated streamflow was indeed significantly higher when the flow was higher for seven out of eight surveys. The exception is the median streamflow for the survey on June 2017 at the Limmat, for which the median estimated streamflow ($Q_{level}$) was not significantly different from the median estimated streamflow during the July and April 2017 surveys, although the first and third quartiles were higher than for the July and April 2017 surveys (see Table 2 and Fig. 8(d)). The variation in streamflow is therefore better represented by streamflow derived from stream level class estimates than by streamflow derived by the factors.

## 4 Discussion

### 4.1 *Can citizens estimate streamflow accurately?*

The results of the streamflow estimation surveys demonstrated the “wisdom of the crowd” effect (Surowiecki 2004, Nielsen 2011) as the median estimates were close to the measured values. However, in practice there will be, at a certain location, only one or at most a few estimates for a certain point in time, so

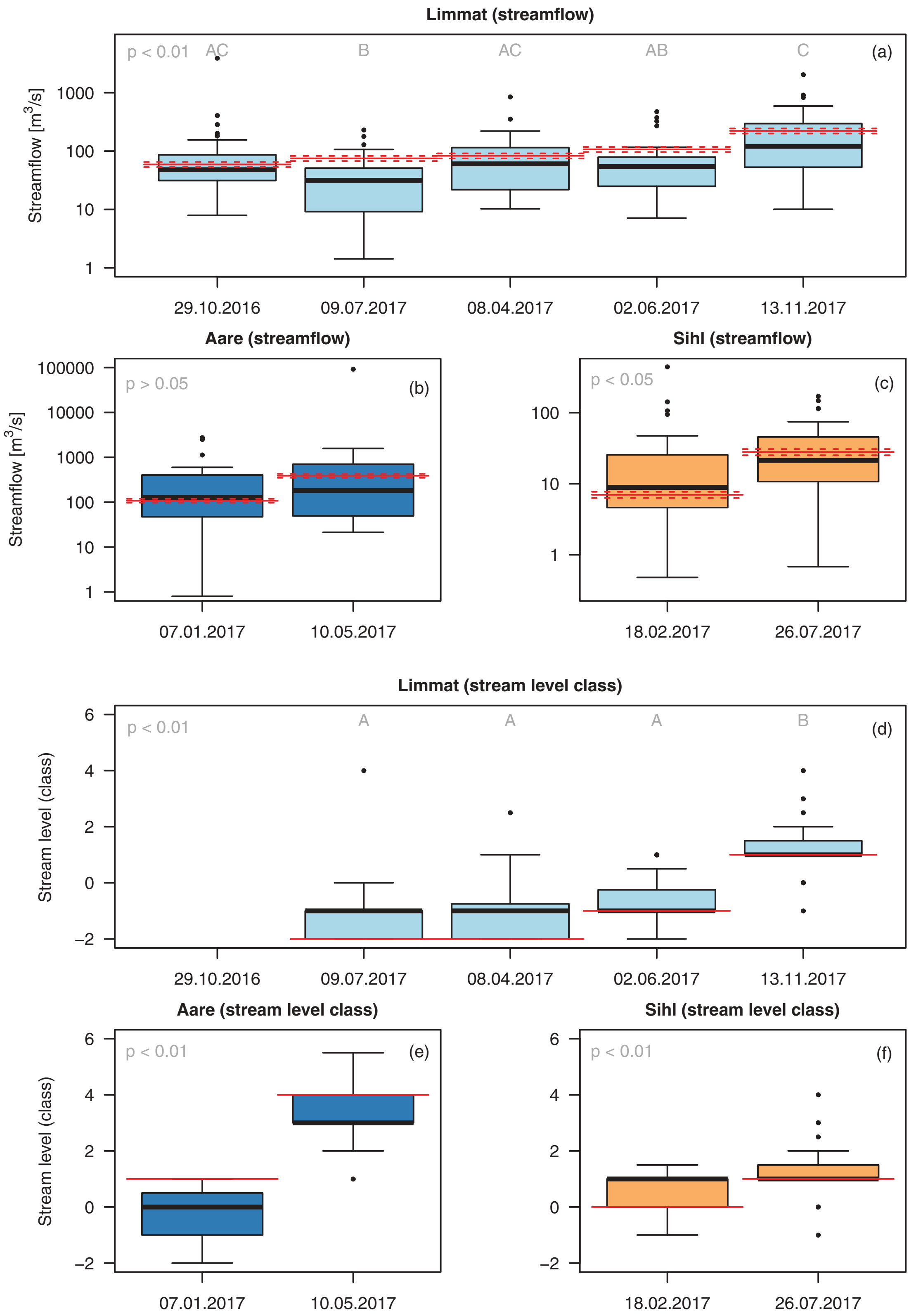



**Figure 8.** Box plots of (a)–(c) the streamflow based on $Q_{factor}$ and (d)–(f) the estimated stream level classes for different flow conditions for three streams (low flow to high flow in each subplot; see Table 1 for details). Solid and dashed (red) lines as described in Figure 6 caption. The red lines indicate the correct values. Note: the axis ranges are different for each stream. The $p$ values indicate the results of the Mann-Whitney (Sihl and Aare) and Kruskal-Wallis (Limmat) tests to determine whether the median estimated streamflow/stream level class of the different surveys are significantly different or not. For the Limmat, surveys with the same upper-case letter (e.g. A) the Dunn *post hoc* test indicated that median streamflow/stream level class estimates were not significantly different from each other.

Table 2. Descriptive statistics of the streamflow derived from the estimated width, mean depth and flow velocity ($Q_{factor}$; $m^3/s$) (and relative estimate, %) and the stream level classes for the Aare, Limmat and Sihl for different flow conditions.

| Stream | Date | Streamflow, $Q_{factor}$ ($m^3/s$) (relative $Q_{factor}$, %) Measured | Percentile 25% | Percentile 50% | Percentile 75% | Stream level class Measured | Percentile 25% | Percentile 50% | Percentile 75% |
|---|---|---|---|---|---|---|---|---|---|
| Sihl | 18.02.2018 | 7 (100) | 5 (66) | 9 (127) | 26 (365) | 0 | 0 | 1 | 1 |
| | 26.07.2018 | 28 (100) | 11 (39) | 21 (76) | 46 (163) | 1 | 2 | 2 | 3 |
| Limmat | 29.10.2016 | 59 (100) | 31 (53) | 48 (81) | 86 (146) | | | | |
| | 08.04.2017 | 83 (100) | 22 (27) | 60 (73) | 111 (134) | −2 | −2 | −1 | −1 |
| | 02.06.2017 | 107 (100) | 26 (24) | 54 (51) | 78 72) | −1 | −1 | −1 | 0 |
| | 09.07.2017 | 75 (100) | 9 (12) | 32 (42) | 49 (66) | −2 | −2 | −1 | −1 |
| | 13.11.2017 | 222 (100) | 53 (24) | 120 (54) | 296 (133) | 1 | 1 | 1 | 2 |
| Aare | 07.01.2017 | 108 (100) | 47 (44) | 128 (118) | 404 (374) | 0 | −1 | 0 | 1 |
| | 10.05.2017 | 389 (100) | 51 (13) | 182 (47) | 684 (176) | 4 | 3 | 3 | 4 |

for hydrological citizen science projects focusing on streamflow the accuracy of the individual estimates is more important than the accuracy of the median estimate.

As expected, estimation of the individual streamflow factors (width, mean depth and flow velocity) led to more accurate streamflow estimates than the direct estimation of streamflow. The reduction in the number of extreme outliers for estimates based on the streamflow factors is likely due to the more intuitive units in which the estimates have to be given. For non-scientists the unit cubic metres per second ($m^3/s$) is difficult to visualize and not easy to relate to everyday experiences. Width and depth in metres (m) and flow velocity in metres per second (m/s) are easier to visualize and estimate for most people. The unit litres per second (L/s) is likely more tangible (as one knows the volume of a litre from drink containers and can estimate how long it takes to fill a bottle or a bucket). This might explain why, for the very small Irchel stream, direct streamflow estimates were more accurate than the streamflow derived from the estimated width, depth and velocity, which included the multiplication of three different types of error. For the Hornbach, another very small stream, there was no significant difference between $Q_{direct}$ and $Q_{factor}$, possibly because it had more streamflow than can fit in a bucket in a second.

The direct streamflow estimates for the Aare (L) were also surprisingly accurate. After the survey, we learned that there used to be a digital display of the current streamflow at the FOEN gauging station, close to the location of our surveys. That display was dismantled before our survey, but it is possible that some participants walked by this site regularly and had a "ballpark" value for the streamflow of the Aare in the back of their minds. Nevertheless, based on our dataset, estimating the streamflow factors rather than the streamflow directly is especially suitable for small and medium streams. It is, however, also important to note that, within the same stream size class, the accuracy of estimates varied for each stream, and even the accuracy of the estimates for the same stream location can vary for different flow conditions (Figs 3 and 8). There was no clear pattern in the relative streamflow estimates ($Q_{factor}$ or $Q_{level}$) to suggest that either low or high flows are more accurately estimated (see Fig. 8 and Table 2; also supplementary Fig. S4).

Many participants estimated the flow velocity fairly accurately if they threw a twig or leaf into the stream, as we suggested, or even just watched something like a bubble in the stream pass by. The differences between these approaches could not be quantified, as it was not documented who chose which approach.

Even though width and mean depth are measured in the same units, width could be estimated more accurately than mean depth. This is consistent with a study by Wahl (1977), in which trained participants measured both the width and depth of a stream, but measured width with more consistency than depth. In our case this is likely due to the refraction of light in water, as well as the inability to see the bottom of the stream because the water is murky or deep, which was the case for the Sihl at high flow (S), Limmat at high flow (M) and both surveys for the Aare (L). Also in some cases – Hornbach (XS), Irchel (XS), Glatt (S), Sihl (S), Töss (S) and Limmat (M) – it was feasible to pace the width along a bridge, in order to gain a better estimate, which made the width estimates more accurate; of course this could not be done for depth. According to Gibson and Bergman (1954), distance estimation can be trained and constant over- and underestimation of distances can be improved.

Training is implemented in many citizen science projects to ensure high-quality data (Bonney *et al.* 2009, Haklay *et al.* 2010, See *et al.* 2013, Stepenuck and Genskow 2017). Participants in our survey received no training, had no prior experience and (presumably) only estimated streamflow and its factors once. The effect of a one-time training was tested for some citizen science projects (Crall *et al.* 2013, Rinderer *et al.* 2015) and has been shown to improve the data-collection ability of the participants. Training options for our study could be in the form of online tutorial videos, or a list of well-known streams and their range in streamflow to indicate approximate numbers for streamflow, as well as width, depth and flow velocity. If participants can improve the accuracy of their estimates and the number of outliers can be reduced sufficiently, streamflow estimates might be usable for hydrological model calibration (Etter *et al.* 2018). Further research will test the applicability of quality control methods, such as outlier detection and the effect of training on the accuracy of streamflow estimates.

The inaccuracies of the streamflow estimates should be seen in light of the rating curve errors that are included in conventional measurements, which have a range of ±20% for medium to high flows and substantially higher errors ranging from −60 to +90% for low flows (McMillan *et al.* 2012). Only 29 and 63% of the $Q_{direct}$ estimates were within ±20 and ±90% of the measured streamflow value, respectively. For the $Q_{factor}$ estimates, the respective values were 15 and 73%.

Ensuring, and possibly improving, the accuracy of the crowdsourced data is an important aspect in any citizen science project. The inaccurate estimates of streamflow might be excluded from analyses by quality control methods. A comprehensive overview of data validation methods in the field of citizen science, such as expert review, photo submission or automatic filtering, is provided by Wiggins *et al.* (2011), and many of these methods are likely also applicable to crowdsourced hydrological estimates.

Video imagery is an alternative way to estimate streamflow. These methods have great potential, especially for more accurately determining flow velocities (Bradley *et al.* 2002, Tsubaki *et al.* 2011, Lüthi *et al.* 2014, Le Coz *et al.* 2016, Tauro *et al.* 2018) and have benefits, such as being more objective and possibly allowing a higher accuracy than visual streamflow estimates. By using advanced and sophisticated technology, they also create a curiosity factor that can motivate people. However, there are also some limitations of these approaches in citizen science projects. Issues include light requirements, camera restrictions and the need for initial *in situ* channel measurements as a reference (Lüthi *et al.* 2014). To encourage more participants to join a citizen science project, we were interested to keep the "installation" of new sites and the observation approach as easy as possible. The visual estimates used in this study are easier to apply for many citizens and, thus, can potentially be used to provide more observations. The different methodologies complement each other and different methods might be most suitable for different locations, participant groups or observation goals. Tauro *et al.* (2018) express a similar opinion: "*Reconciling and complementing observations from such an abundant pool of methodologies, devices and platforms is the ultimate goal of the research community towards an improved understanding of hydrological processes*" (Tauro *et al.* 2018, p. 187). Many of the current limitations in video imagery will likely be resolved in the future, making this approach a more usable alternative for streamflow or stream level estimates. A possibility in the future might also be to develop a virtual staff gauge in an augmented reality setting, thereby facilitating participants' stream level class estimates.

### 4.2 Can citizens estimate stream level classes accurately?

Stream level classes were introduced to simplify the stream level estimation task for the participants. In theory we could have also asked participants to estimate a metric value above or below some fixed point. However, the depth estimates (Fig. 4) for $Q_{factor}$ suggest that this approach would lead to estimates with a low accuracy. The high accuracy of stream level class estimates and the small number of outliers (i.e., estimates that are more than one class off target) indicate that this is a suitable parameter for citizen science projects. The major benefits of the virtual staff gauge approach is that estimates can be done quickly and that relative variations in stream level can be estimated with small uncertainties, but, on the down side, they also have a lower resolution. A participant can be no more than 10 classes off target (which never happened; 0.7% of participants were four classes off and <0.5% of participants were five or six classes off).

Participants only needed to compare the current stream level to a previous stream level using structures, streambanks or stones as a reference. If the virtual staff gauge is well placed (i.e., there is a suitable structure on the stream bank or in the stream), the participant only needs to look for the reference and then determines the corresponding stream level class. In general, the vast majority of participants had no problem understanding the concept and estimated the stream level class correctly; outliers in the estimated stream level classes were very rare. However, there were also a few clearly wrong stream level class estimates, which might suggest a misunderstanding of the concept by some participants. The two most extreme overestimations were both at the Limmat, the most extreme underestimations at the Aare. Most participants (49%) underestimated the stream level class at the Aare. The reasons are unknown, but potentially this could be attributed to a staff gauge placement during an exceptionally low stream level (less than a 2-year low according to official measurements; BAFU 2017), meaning that the zero value was already very low. This might have confused participants as they may have thought that the staff gauge represents the average streamflow condition.

The stream level class estimates were especially accurate for smaller streams where the opposite stream banks, at which the virtual staff gauges were located in the photo, were close to the participant. The Limmat is a wider stream, but was an exception as the virtual staff gauge was placed on a bridge pillar, which was relatively close to the observer. This is most likely the reason why the stream level class estimates for the Limmat were more accurate than for the Aare (the only stream where the references for the virtual staff gauge were 50–60 m away from the participant), even though the widths of the actual streams were similar (50 and 52 m, respectively). This shows that, for stream level class estimates, the placement of the virtual staff gauge is important. One of the very small streams (Irchel) had a poorly placed staff gauge (the image was taken looking down onto the stream rather than horizontally from the height of the stream level, which distorted the virtual staff gauge relative to the wall behind the stream) and made it more difficult to read. The median relative estimate for $Q_{level}$ for the Irchel stream was 12%, whereas the median relative estimate for $Q_{level}$ for all surveys was 101%.

Several studies have examined the accuracy of crowdsourced data (Haklay *et al.* 2010, Crall *et al.* 2011, See *et al.* 2013, Isaac and Pocock 2015, Tye *et al.* 2016, Aceves-Bueno *et al.* 2017, Mengersen *et al.* 2017), mentioning case studies such as OpenStreetMaps, where Volunteered Geographic Information (VGI) data are collected online and verified by other participants (Haklay *et al.* 2010), and discussing issues such as presence-only data for crowdsourced species classification (Isaac and Pocock 2015, Tye *et al.* 2016, Mengersen *et al.* 2017). While hydrological studies have also discussed crowdsourced data accuracy (Turner and Richter 2011, Rinderer *et al.* 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer *et al.* 2015, Le Coz *et al.* 2016, Little *et al.* 2016, Weeser *et al.* 2018), most of these studies looked at crowdsourced measurements rather than estimates (Lowry and Fienen 2013, Peckenham and Peckenham 2014, Little *et al.* 2016, Weeser *et al.* 2018). While others, such as Turner and Richter (2011), looked at class estimates, they mainly looked at two class options (wet or dry stream), but unfortunately do not mention data accuracy apart from the fact that participants were trained for consistency. Rinderer *et al.* (2012, 2015), who also looked at classed data, analysed participants' ability to estimate relative soil moisture classes and found that, in one case study, 95% of participants were no more than one class off (Rinderer *et al.* 2012), and in another study with various groups, 81–93% of the participants were no more than one class off (Rinderer *et al.* 2015). However, as far as we are aware, our study is the first to address the accuracy of participants' estimates of stream level classes.

In addition to being more accurate, the stream level class estimation process is also very quick, which is a big advantage for a citizen science project. It is assumed that offering a fast procedure to document stream levels will encourage citizen observers to contribute data to a project regularly (Eveleigh *et al.* 2014). It is very common for citizen science projects that the majority of the contributions come from a small group of high contributors (Lowry and Fienen 2013, Eveleigh *et al.* 2014, Sauermann and Franzoni 2015). For example, in the CrowdHydrology project, one participant walked past a particular station three to four times a week, which led to this station having almost 10 times as many measurements as the station with the next highest number of data submissions (Lowry and Fienen 2013). This highlights the extreme value of these high contributors and shows that it is important to be able to take measurements quickly.

### 4.3 Are citizens likely to observe variations in streamflow?

Having data for high and low flows, or relative variations in streamflow is crucial in order to determine how a stream reacts to precipitation, snowmelt events or long periods without rainfall, and for hydrological model calibration. Hence, it is important to know if crowdsourced data can properly reflect such variations in streamflow and whether the accuracy of the data depends on the flow conditions. The results from the surveys suggest that the temporal dynamics in streamflow will be relatively poorly represented by citizen-based streamflow estimates. For two of the three streams (Sihl and Aare), the median streamflow was overestimated at low flows and underestimated at high flows, which indicates insufficient adjustment of the streamflow estimates to the variation in flow conditions. For the Limmat, the significant difference in the streamflow estimates does not seem to correspond to the differences in the measured streamflow (Fig. 8(a)–(c)). This is partly due to the problem that width (and to a lesser degree velocity) estimates were more accurate compared to depth estimates (Fig. 4). As long as a high flow stays within the streambank, the width of the streams in our survey does not vary significantly between low and high flows. Thus, the majority of the variation in flow conditions is due to the variation in depth, which was most difficult to estimate.

During the surveys we did not ask the same persons to estimate the flow during high and low flow conditions. The results for an individual who reports the streamflow at different times may be different, because the participant might consistently over- or underestimate the flow and therefore the relative variations might be more accurate than indicated by our results (Rinderer *et al.* 2015). Thus, further research is needed to determine if the streamflow dynamics are better described by the streamflow estimates when the majority of the contributions for a particular stream are made by one (or a few) active citizen(s) (Lowry and Fienen 2013).

The high and low flow patterns are better reflected in the stream level class estimates, with the median flow derived from these estimates ($Q_{level}$) being significantly different between high and low flows for all streams. For the Limmat, the *post hoc* tests showed a significant difference between the high flow and all other survey campaign estimates. This underlines the benefits of collecting stream level class estimates, particularly for model calibration (see additional discussion below).

### 4.4 Should citizen science projects focus on streamflow or stream level class estimates?

The reduction of the number of outliers in the streamflow estimates calculated from the stream level class data ($Q_{level}$) compared to the direct streamflow estimates ($Q_{direct}$) and streamflow estimates based on the streamflow factors ($Q_{factor}$) can partly be explained by the limited number of potential entries for the virtual staff gauge (i.e., participants can only choose one out of 10 available classes for the stream level estimate). For $Q_{direct}$ and $Q_{factor}$, participants were able to state any value for their estimates, even values that are physically impossible for a particular stream. Hence, with regard to the reduction of outliers, estimating stream level classes seems advantageous for citizen science projects. Additionally, our results suggest that stream level class estimates appear to be better suited to represent variations in flow conditions. Thus, the results of this study suggest that citizen science projects should focus on stream level class estimates instead of streamflow estimates, although this needs to be tested for different climatic, geographical and socio-economic settings.

However, it should be noted that part of the difference in accuracy for the stream level class estimates and streamflow estimates is due to the difference between relative and absolute values. For our approach, it would be impractical to use classes for streamflow estimates, as we would need many classes, or the resolution of the data would be very low (i.e., the flow for a given stream is likely to always be within the same class). However, as mentioned above, lists of well-known streams, giving their streamflow range to indicate orders of magnitude for the expected streamflow, as well as width, depth and flow velocity, could be provided to make it easier for citizens to make the estimates and to improve the accuracy of the estimates.

One of the disadvantages of the stream level classes is that each class represents a range of potential streamflow values, rather than one specific value. If a participant estimates that the stream level is in class two, it is unclear whether that means the upper, middle or lower part of the class. The other disadvantage is that these estimates do not provide information on streamflow volumes. However, the usability of stream level class data for hydrological model calibration was tested by van Meerveld *et al.* (2017), who showed that stream level class data can be used to calibrate a simple bucket-type hydrological model, and suggested that simple hydrological models can be used to convert stream level class data to time series of streamflow. The value of stream level data for hydrological model calibration, especially for humid catchments, was demonstrated recently by Seibert and Vis (2016). The value of crowdsourced stream level data (photographs of a fixed staff gauge) together with rainfall and flood observations was also shown by Starkey *et al.* (2017). They used community-based observations of rainfall (manual raingauges), river levels (manual staff gauge) and flood-related evidence (anecdotes, photographs or videos) alongside traditional information (tipping bucket raingauge, official raingauge measurements, six pressure transducers for water level measurements and flow gauging for the discharge-rating curve), in order to fill spatial and temporal gaps in hydrometric data for a 42 km$^2$ catchment in the UK to improve a physically-based, spatially-distributed catchment model (SHETRAN). Etter *et al.* (2018) calibrated a bucket-type model with synthetic crowdsourced streamflow data with different degrees of error (including errors that are comparable to those observed in this study) and different temporal resolutions, and indeed found that such streamflow estimates do not contain sufficient information to improve the model compared to random parameter sets. However, they also showed that, if the standard deviation of the log-normal distribution that was used to describe the errors of crowdsourced streamflow estimates could be reduced by a factor of two, one estimate per week would lead to a significant improvement in the model simulations.

## 5 Conclusion

We asked 517 citizens to estimate streamflow directly and indirectly by estimating the stream width, depth and flow velocity. We also asked them to estimate the stream level class. The survey results allowed us to quantify the accuracy of the estimates and are, thus, a basis for evaluating the potential value of citizen science based estimates of streamflow and stream level classes. The median estimated streamflow values were close to the measured streamflow, but there were also many outliers, and the variations in the flow conditions were not fully discernible in the streamflow estimates. The stream level class estimates, which were converted into streamflow values for comparison, had far fewer outliers and were significantly different for the different flow conditions. Stream level class estimates also seemed to be quicker and easier to estimate and are thus considered preferable for citizen science approaches. Hydrological models can then be parameterized based on these stream level class estimates to obtain streamflow time series. The study was conducted in Switzerland and, while we do not expect significant differences, we recommend testing the accuracy of citizen science based estimates of streamflow and stream level classes in different climatic, geographical or socio-economic settings and for rivers with different sizes.

## Acknowledgements

We thank all study participants for their time and interest in this research project and for sharing their hydrological estimates with us, as well as the FOEN (Federal Office for the Environment) and WWEA (Office of Waste, Water, Energy and Air of Canton Zurich) for providing the streamflow data used for comparison with the estimates.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This study was funded by the Swiss National Science Foundation (project 163008, CrowdWater) [Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung 163008, CrowdWater].

## ORCID

Barbara Strobl http://orcid.org/0000-0001-5530-4632
Simon Etter http://orcid.org/0000-0002-7553-9102
Ilja van Meerveld http://orcid.org/0000-0002-7547-3270
Jan Seibert http://orcid.org/0000-0002-6314-2124

## References

Aceves-Bueno, E., *et al.*, 2017. The accuracy of citizen science data: a quantitative review. *The Bulletin of the Ecological Society of America*, 98 (4), 278–290. doi:10.1002/bes2.1336

BAFU, 2017. *Niedrigwasserwahrscheinlichkeit (Jahresniedrigwasser NM7Q) Aare-Brugg (EDV: 2016)*. Available from: https://www.hydrodaten.admin.ch/lhg/sdi/nq_studien/nq_statistics/2016nq.pdf [Accessed 2 May 2018].

Beven, K. and Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, 25 (10), 1676–1680. doi:10.1002/hyp.v25.10

Beven, K.J., 2012. *Rainfall-runoff modelling: the primer*. 2nd ed. Oxford, UK: Wiley-Blackwell.

Bishop, K., *et al.*, 2008. Aqua Incognita: the unknown headwaters. *Hydrological Processes*, 22, 1239–1242. doi:10.1002/hyp.7049

Bonney, R., *et al.*, 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59 (11), 977–984. doi:10.1525/bio.2009.59.11.9

Bradley, A.A., *et al.*, 2002. Flow measurement in streams using video imagery. *Water Resources Research*, 38 (12). doi:10.1029/2002WR001317

Breuer, L., *et al.*, 2015. HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters. *Scientific Reports*, 5, 16503. doi:10.1038/srep16503

Buytaert, W., *et al.*, 2014. Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, 2 (26), 21. Available from: http://journal.frontiersin.org/article/10.3389/feart.2014.00026/abstract

Crall, A.W., *et al.*, 2011. Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, 4 (6), 433–442. doi:10.1111/j.1755-263X.2011.00196.x
Crall, A.W., *et al.*, 2013. The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy. *Public Understanding of Science*, 22 (6), 745–764. doi:10.1177/0963662511434894
Dickinson, J.L., Zuckerberg, B., and Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41 (1), 149–172. doi:10.1146/annurev-ecolsys-102209-144636
Dingman, S.L., 2015. *Physical Hydrology*. 3rd ed. Long Grove: Waveland Press Inc.
Dunn, O.J., 1964. Multiple Comparisons Using Rank Sums. *Technometrics*, 6 (3), 241–252. doi:10.1080/00401706.1964.10490181
Engel, S.R. and Voshell, J.R., 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? *American Entomologist*, 48, 164–177. doi:10.1093/ae/48.3.164
Etter, S., *et al.*, 2018. Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, 22, 5243–5257. doi:10.5194/hess-22-5243-2018
Eveleigh, A., *et al.*, 2014. Designing for dabblers and deterring drop-outs in citizen science. *In*: *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 26 April–1 May, Toronto. New York, NY: Association for Computing Machinery, 2985–2994. doi:10.1145/2556288.2557262
Field, A., Miles, J., and Field, Z., 2013. *Discovering statistics using R*. Los Angeles: Sage.
Gibson, E.J. and Bergman, R., 1954. The effect of training on absolute estimation of distance over the ground. *Journal of Experimental Psychology*, 48 (6), 473–482. doi:10.1037/h0055007
Hadj-Hammou, J., *et al.*, 2017. Getting the full picture: Assessing the complementarity of citizen science and agency monitoring data. *PLoS ONE*, 12 (12), 1–18. doi:10.1371/journal.pone.0188507
Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37 (4), 682–703. doi:10.1068/b35097
Haklay, M., (Muki), *et al.*, 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47 (4), 315–322. doi:10.1179/000870410X12911304958827
Herschy, R.W., 1971. *The magnitude of errors at flow measurement stations*. Technical report, Water Resources Board, Reading, UK.
Isaac, N.J.B. and Pocock, M.J.O., 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115 (3), 522–531. doi:10.1111/bij.12532
Juston, J., Seibert, J., and Johansson, P., 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, 23 (21), 3093–3109. doi:10.1002/hyp.7421
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 41, W03S04. doi:10.1029/2005WR004362
Kundzewicz, Z.W., 1997. Water resources for sustainable development. *Hydrological Sciences Journal*, 42 (4), 467–480. doi:10.1080/02626669709492047
Le Coz, J., *et al.*, 2016. Crowdsourced data for flood hydrology: feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology*, 541, 766–777. doi:10.1016/j.jhydrol.2016.07.036
Little, K.E., Hayashi, M., and Liang, S., 2016. Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach. *Groundwater*, 54 (3), 317–324. doi:10.1111/gwat.2016.54.issue-3
Lowry, C.S. and Fienen, M.N., 2013. CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, 51 (1), 151–156. doi:10.1111/j.1745-6584.2012.00956.x
Lüthi, B., Philippe, T., and Peña-Haro, S., 2014. Mobile device app for small open-channel flow measurement. *In*: D.P. Ames, N.W.T. Quinn, and A.E. Rizzoli, eds. *7th International Congress on Environmental Modelling and Software*. San Diego, CA. Available from: http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_112.pdf
Manning, R., 1891. On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, 20, 161–207.
McMillan, H., Krueger, T., and Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, 26 (26), 4078–4111. doi:10.1002/hyp.v26.26
Mengersen, K., *et al.*, 2017. Modelling imperfect presence data obtained by citizen science. *Environmetrics*, 28 (5), 1–29. doi:10.1002/env.2446
Nielsen, M., 2011. *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.
Peckenham, J.M. and Peckenham, S.K., 2014. Assessment of quality for middle level and high school student-generated water quality data. *Journal of the American Water Resources Association*, 50 (6), 1477–1487. doi:10.1111/jawr.12213
Pelletier, P.M., 1988. Uncertainties in the single determination of river discharge: a literature review. *Canadian Journal of Civil Engineering*, 15 (5), 834–850. Available from: http://www.nrcresearchpress.com/doi/10.1139/l88-109
Perrin, C., *et al.*, 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, 52 (1), 131–151. Available from: http://www.tandfonline.com/doi/abs/10.1623/hysj.52.1.131
Pool, S., Viviroli, D., and Seibert, J., 2017. Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, 554, 613–622. doi:10.1016/j.jhydrol.2017.09.037
Rinderer, M., *et al.*, 2012. Sensing with boots and trousers - qualitative field observations of shallow soil moisture

patterns. *Hydrological Processes*, 26 (26), 4112–4120. Available from: 10.1002/hyp.9531 [Accessed 27 Mar 2014].

Rinderer, M., *et al.*, 2015. Qualitative soil moisture assessment in semi-arid Africa - the role of experience and training on inter-rater reliability. *Hydrology and Earth System Sciences*, 19, 3505–3516. doi:10.5194/hess-19-3505-2015

Ruhi, A., Messager, M.L., and Olden, J.D., 2018. Tracking the pulse of the Earth's fresh waters. *Nature Sustainability*, 1 (4), 198–203. doi:10.1038/s41893-018-0047-7

Sauermann, H. and Franzoni, C., 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112 (3), 679–684. doi:10.1073/pnas.1408907112

See, L., *et al.*, 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PlosONE*, 8 (7), 1–11. doi:10.1371/journal.pone.0069958

Seibert, J. and Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, 13 (6), 883–892. Available from: http://www.hydrol-earth-syst-sci.net/13/883/2009/

Seibert, J. and McDonnell, J.J., 2015. Gauging the ungauged basin: relative value of soft and hard data. *Journal of Hydrologic Engineering*, 20 (1), A4014004-1–6. Available from: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000861

Seibert, J. and Vis, M.J.P., 2016. How informative are stream level observations in different geographic regions? *Hydrological Processes*, 30 (14), 2498–2508. doi:10.1002/hyp.10887

Starkey, E., *et al.*, 2017. Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. *Journal of Hydrology*, 548, 801–817. doi:10.1016/j.jhydrol.2017.03.019

Statistik Stadt Zürich, 2017. Statistisches Jahrbuch der Stadt Zürich 2017, 188–201. Available from: https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/Jahrbuch/statistisches-jahrbuch-der-stadt-zuerich_2017.html [Accessed 12 Oct 2017].

Stepenuck, K.F. and Genskow, K.D., 2017. Characterizing the breadth and depth of volunteer water monitoring programs in the united states. *Environmental Management*, 61 (1), 46–57. doi:10.1007/s00267-017-0956-7

Surowiecki, J., 2004. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. London, UK: Little Brown.

Tauro, F., *et al.*, 2018. Measurements and observations in the XXI century (MOXXI): Innovation and multi-disciplinarity to sense the hydrological cycle. *Hydrological Sciences Journal*, 63 (2), 169–196. doi:10.1080/02626667.2017.1420191

Tsubaki, R., Fujita, I., and Tsutsumi, S., 2011. Measurement of the flood discharge of a small-sized river using an existing digital video recording system. *Journal of Hydro-Environment Research*, 5 (4), 313–321. doi:10.1016/j.jher.2010.12.004

Tulloch, A.I.T., *et al.*, 2013. Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165, 128–138. doi:10.1016/j.biocon.2013.05.025

Turner, D.S. and Richter, H.E., 2011. Wet/dry mapping: using citizen scientists to monitor the extent of perennial surface flow in dryland regions. *Environmental Management*, 47 (3), 497–505. doi:10.1007/s00267-010-9607-y

Tye, C.A., *et al.*, 2016. Evaluating citizen versus professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, 54 (2), 628–637.

van Meerveld, H.J., Vis, M.J.P., and Seibert, J., 2017. Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, 21 (9), 4895–4905. doi:10.5194/hess-21-4895-2017

Vis, M., *et al.*, 2015. Model calibration criteria for estimating ecological flow characteristics. *Water (Switzerland)*, 7 (5), 2358–2381.

Wahl, K.L., 1977. Accuarcy of channel measurements and the implications in estimating streamflow characteristics. *Journal Research of the U.S. Geological Survey*, 5 (6), 811–814.

Weeser, B., *et al.*, 2018. Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of The Total Environment*, 632, 1590–1599. doi:10.1016/j.scitotenv.2018.03.130

Welber, M., *et al.*, 2016. Field assessment of noncontact stream gauging using portable surface velocity radats (SVR). *Water Resources Research*, 52, 1108–1126. doi:10.1002/2015WR017906

Westerberg, I., *et al.*, 2011. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25 (4), 603–613. doi:10.1002/hyp.7848

Wiggins, A., *et al.*, 2011. Mechanisms for data quality and validation in citizen science. *In*: *Seventh IEEE International Conference on e-Science Workshops*, 5–8 December. Stockholm: IEEE, 14–19. doi:10.1109/eScienceW.2011.27

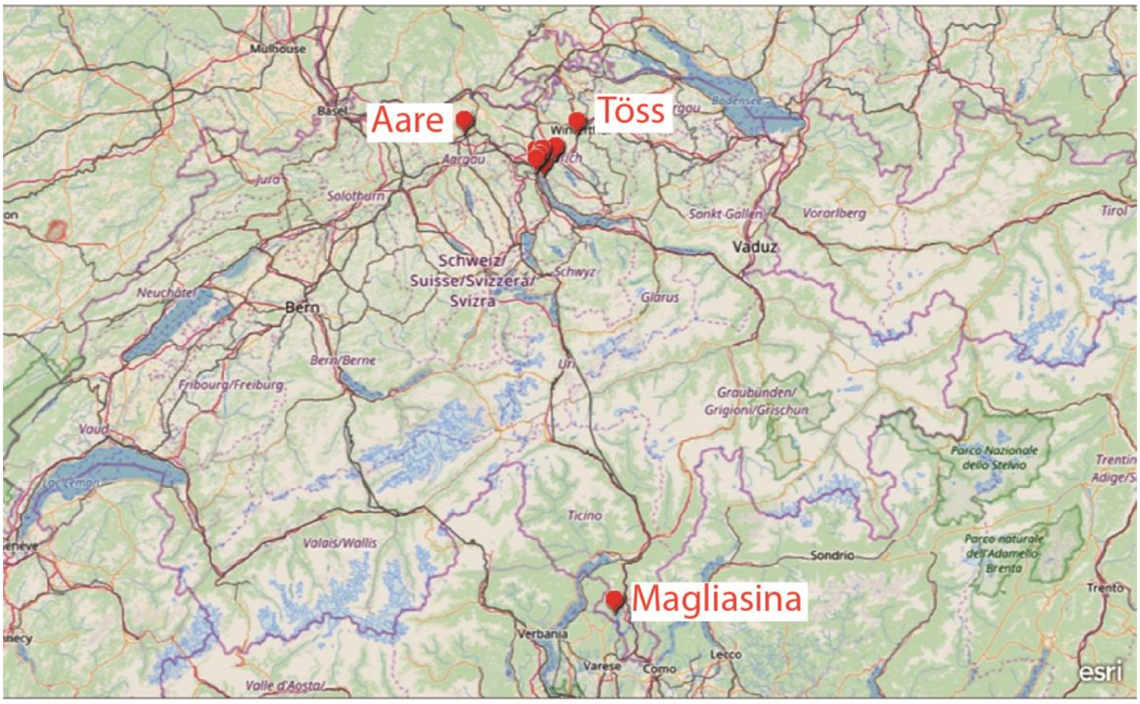
## Supplementary material

## Accuracy of crowdsourced streamflow and stream level class estimates

Barbara Strobl[a]*, Simon Etter[a], Ilja van Meerveld[a], and Jan Seibert[a,b]

*[a] Department of Geography, University of Zurich, Zurich, Switzerland; [b] Department of Earth Sciences, Uppsala University, Uppsala, Sweden*

**S1 Map with the survey locations**





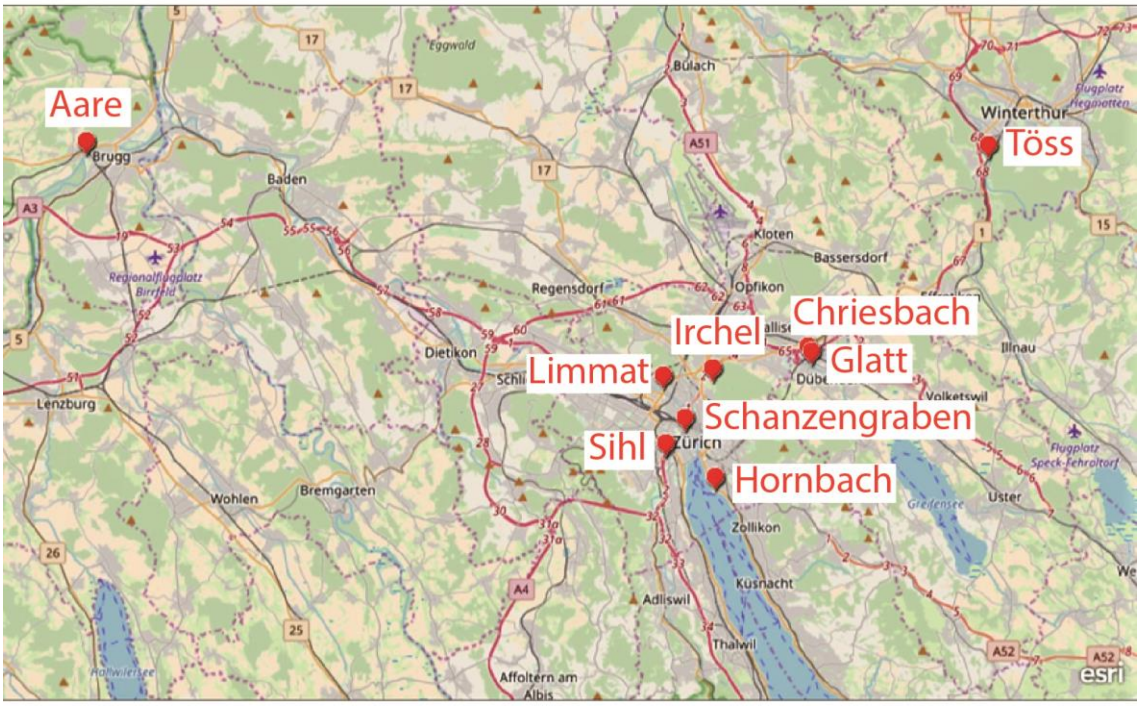






**Figure S1.** (a) Map of Switzerland showing the location of all 10 survey locations and (b) map of the greater Zurich area, showing the location of the nine field surveys around Zurich. For details of the surveys, see Table 1 in the main article. Background map from OpenStreetMap.

*S2 Example of the forms used for the surveys (Limmat)*

Limmat

crowd water

## Water Level & Streamflow

The fields marked with * are required.

**Date***:______________________ **Time***:______________________

Age: ________________________

Gender:
- Female
- Male
- Other

Mother tongue: ________________

Highest level of education:
- Secondary
- Apprenticeship
- Pre-university (A-Levels)
- University/ Applied University

Have you completed this form before? o yes o no
If so, how often? _________________

1. Water Level


In which water level category of the virtual scale on the picture would the current water level in the river be?

Category*: ____________

2. Streamflow

How high is the streamflow in this river in $m^3/s$?

Streamflow [$m^3/s$]*:

________________________________________

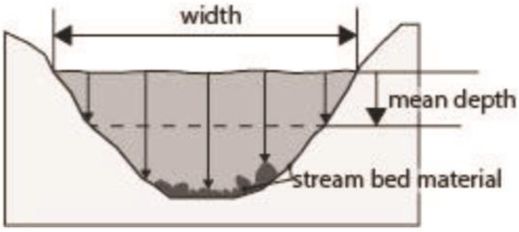
PLEASE TURN THE PAGE!

In order to estimate the runoff better, you need values for the width, the average depth and the flow velocity. This estimate can be further improved if the bed material type is also specified.

Width [m]*: ______________________________

Mean depth [m]*: ______________________________

Material type in river bed*: ______________________________



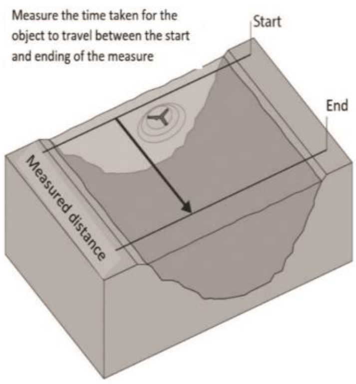

Decide for the average rock size.

TIPP: Choose between:

- **Sand**
- **Gravel** (Smaller than a chicken egg)
- **Cobbles** (Bigger than a chicken egg but less than 20cm diameter)
- **Boulders** (over 20cm diameter)
- **Solid bedrock**.

Flow velocity [m/s]:

- Distance [m]: ______________________________
- Time [s]: ______________________________




TIPP: Throw a twig into the river and stop the seconds the twig needs to flow down a predetermined distance, e.g. 3 meters. Flow velocity can be calculated by dividing the meters by the seconds. If you do not want to calculate, just write the distance and time data separately.

What would motivate you to make these estimates independently with an app?

______________________________________________________________

______________________________________________________________

## Thank you for your contribution to our research!

*S3 Participant demographic*

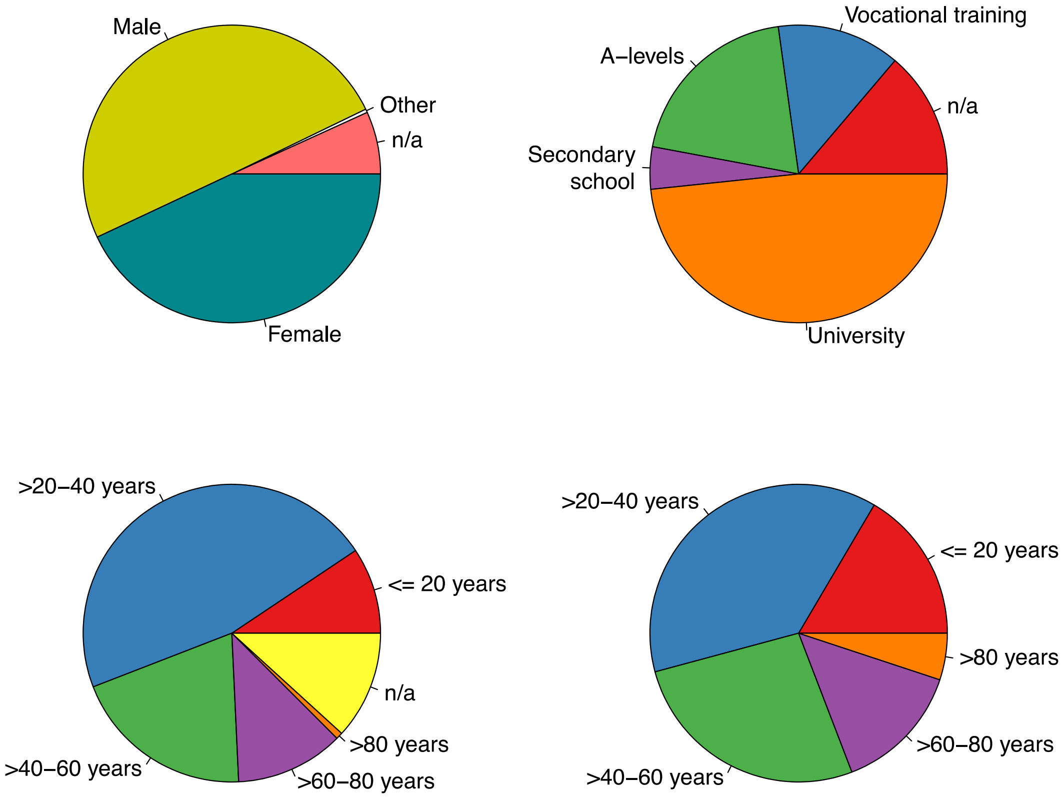



**Figure S2.** (a) Gender, (b) education and (c) age distribution of the participants, and (d) age distribution in the city of Zurich for comparison *(Data source: Statistik Stadt Zürich 2017)*.

*S4 Relative stream level class estimates for small sized streams*

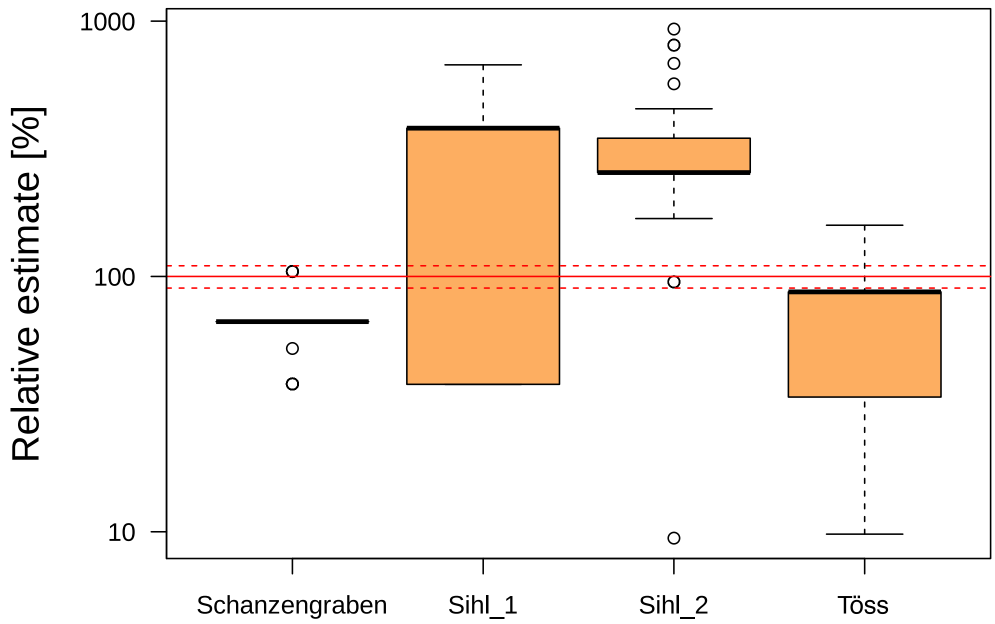



**Figure S3.** Relative stream level class estimates for small streams, converted into streamflow using the midpoint of each level class for each estimate. Red lines indicate the measured streamflow and the dashed red line indicates the 10% uncertainty associated with the measured streamflow. The boxplot shows the high variability in the estimates for Sihl_1 and Töss.

## *S5 Relative streamflow estimates during high and low flow*

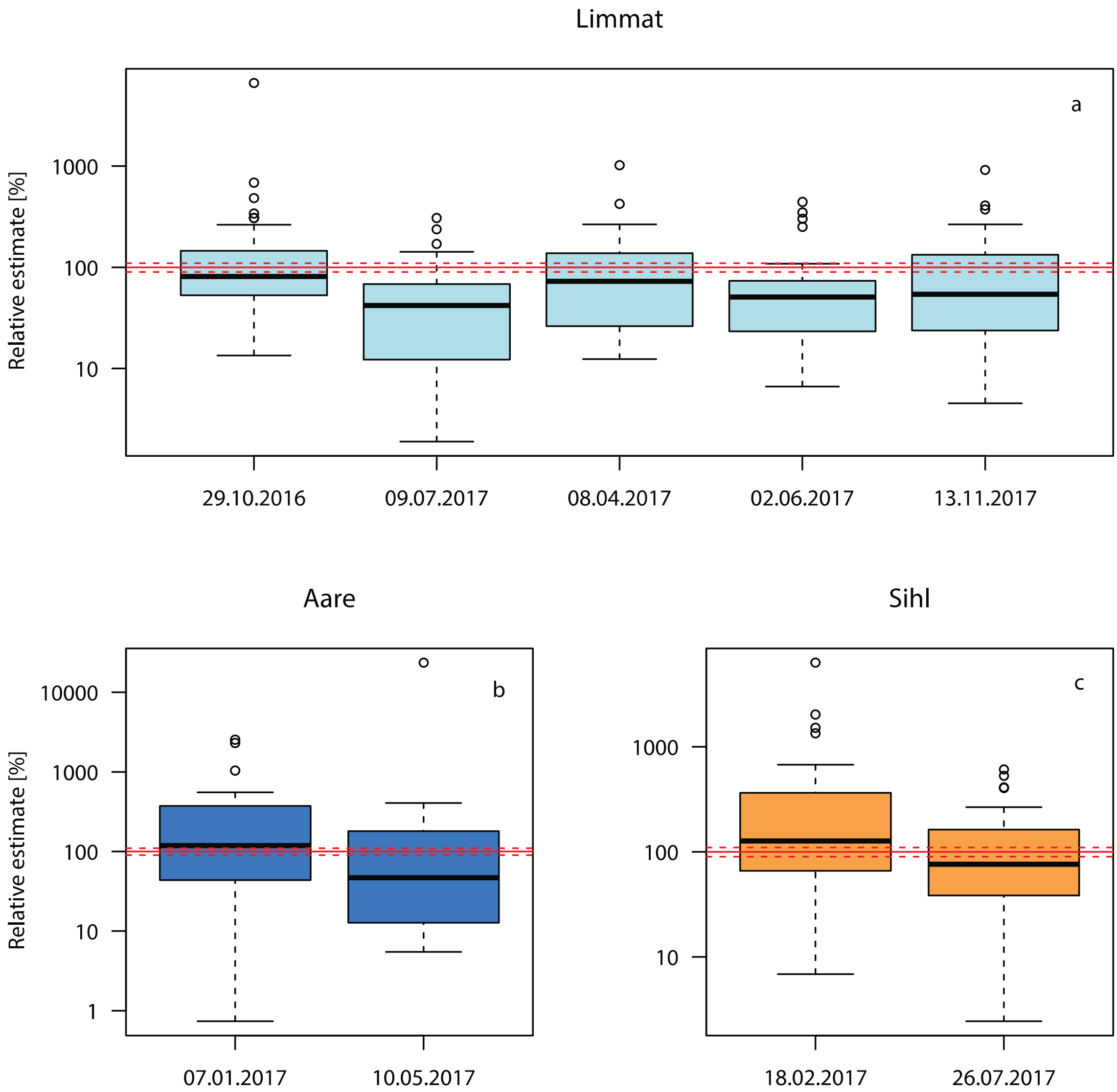



**Figure S4.** Boxplots of the relative estimates of streamflow based on the estimated width, mean depth and flow velocity ($Q_{factor}$) for surveys under different flow conditions at the Limmat, Aare and Sihl. Red lines indicate the measured streamflow and the dashed red line indicates the 10% uncertainty associated with the measured streamflow. For details on the flow conditions during the surveys, see Table 1 in the main article.

# Paper IV

To be resubmitted to Hydrological Processes after moderate revisions before May 31, 2020.

# Quality and timing of crowd-based water level class observations

Simon Etter[1], Barbara Strobl[1], Ilja van Meerveld[1], Jan Seibert[1,2]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, 75007 Uppsala, Sweden.

Corresponding author: Simon Etter, simon.etter@outlook.com

Keywords: Hydrology, citizen science, crowdwater, water level class, smartphone

Funding: Swiss National Science Foundation (project no. 163008).

## Key Points

- Changes in stream water levels observed by citizens based on virtual staff gauges agreed well with measured changes in water levels
- Observation uncertainties depended mainly on the placement of the virtual staff gauge
- Data collected by individual observers using a smartphone app were of higher quality than those collected by multiple observers using paper forms

## Abstract

Crowd-based hydrological observations can supplement existing monitoring networks and allow data collection in regions where otherwise no data would be available. In the citizen science project CrowdWater, repeated water level observations using a virtual staff gauge approach result in time series of water level classes. To investigate the quality of these observations, we compared the water level class data from nine locations where citizen scientists reported multiple observations using a smartphone app and at twelve other locations where signposts were set up to ask citizens to record observations on a form that could be left in a letterbox, with the nearest measured water levels from the same stream. The results indicate that the quality of the data collected with the app was higher than for the forms. A possible explanation is that for each app location, most contributions were made by a single person, whereas at the locations of the forms almost every observation was made by a new contributor. On average, more contributions were made between May and September than during the other months. Observations were submitted for a range of flow conditions, with a higher fraction of high flow observations for the data collected with the app. Overall, the results are encouraging for citizen science approaches in hydrology and demonstrate that the smartphone application with its virtual staff gauge is a promising approach for crowd-based water level class observations.

# 1 Introduction

Hydrometric networks provide basic information for water management (Mishra and Coulibaly, 2009). However, in many regions of the world, the hydrological measurement infrastructure is limited or poorly maintained (Hannah et al., 2011; Sivapalan, 2003). These areas often coincide with areas that are vulnerable to extreme conditions and events (Walker et al., 2016) and where data would thus be highly beneficial. One possibility to overcome this data limitation is to involve the public in hydrological observations using citizen science approaches. Citizen science can provide data at many more locations than official agencies are able to do, and thereby can complement the data from official monitoring networks. Examples are the citizen observatories WeSenseIt (www.wesenseit.com; Lanfranchi et al., 2014), GroundTruth2.0 (https://gt20.eu) and SCENT (https://scent-project.eu). Citizen science projects can potentially also collect data in regions where otherwise no data are available to allow calibration of models, data-based measures for protection, or warning systems against water-related natural hazards. Some of the existing examples of citizen science projects that collect streamflow or water level data for ungauged streams are CrowdHydrology (Lowry et al., 2019), a project in Kenya (Weeser et al., 2018), CitHyd (www.cithyd.com, Balbo & Galimberti, 2016) in Italy, SmartPhones4Water in Nepal (www.smartphones4water.org; Davids et al., 2017) and CrowdWater (www.crowdwater.ch; Seibert et al., 2019). The CrowdWater project uses a smartphone application (hereafter referred to as "app") to collect information on water level changes in streams using a virtual staff gauge (Seibert et al., 2019). As a first test, Strobl et al. (2019a) asked passers-by at ten river locations in Switzerland to estimate both the streamflow and the water level class (hereafter shortened to WL-class) based on the virtual staff gauge approach and quantified the errors of these estimates. These errors were then used to create synthetic streamflow and WL-class time series in two model studies to explore the potential value of such data for model calibration (Etter et al., 2018; 2020). The studies showed that the estimates of streamflow were not accurate enough to be informative for hydrological model calibration but that WL-class estimates significantly improved model performance compared to the situation without any data. The study assumed one observation per week on average for calibration, which resulted in simulations, that were almost as good as those obtained using continuous water level data (i.e., data that could be obtained from a water level logger).

In this study, we evaluate the quality of WL-class data collected at real CrowdWater locations. Between April 2017 and September 2019, more than 4475 WL-class observations were made with the app (Figure S1). These observations were made at more than 816 locations, including 26 locations with more than 30 repeated observations. The accuracy of these data may be different from the previous study (Strobl et al., 2019a), where the experts were physically present. The data were collected over a one-year period or more (rather than one day) and, thus, cover a much wider range in water levels. In other words, the data analysed in this study are real data that were collected by citizen scientists in the CrowdWater project. We used data from nine locations where data were collected with the CrowdWater app and twelve locations where observations were collected using paper forms and letterboxes. There was a wider range in the way that the virtual staff gauges were set-up because all app spots (except A5) were initiated by real citizen scientists; the reference images with the virtual staff gauges at the pen-and-paper locations were created by ourselves. For all locations, measured water level data were available from either the same location or a nearby site on the same stream. Furthermore, we analysed when the observations were submitted to see whether there is temporal bias in these data (e.g., whether observations are only made on certain days or only during low flow periods or cover the entire range of conditions).

# 2 Methods

## 2.1 Virtual staff gauge approach

The CrowdWater project started in April 2016 and the app was released in April 2017. As of September 2019, about 373 different citizen scientists had reported 4475 WL-class observations with the app (Figure 1). Citizen scientists can either start their own time series of water level observations or contribute to the time series at an existing observation location (hereafter referred to as "spot", because in the app they are called spots). All existing spots are displayed on a map in the app (*Figure 1a*). For each new spot a photograph of the stream is taken perpendicularly to the flow direction. The citizen scientist then inserts a virtual staff gauge with ten classes onto the picture. The size of the staff gauge can be adjusted to the size of the stream, and the staff gauge needs to be moved so that the class zero is aligned with the  water level in the picture (Seibert et al., 2019). Subsequent observations of the WL-class are made with the help of the virtual staff gauge by comparing the current water level with the virtual staff gauge in this reference picture using the features on the opposite side of the stream, or bridge pillars and stones as a reference (Figure 1, Seibert et al., 2019).

[Figure 1 here]

## 2.2 Study locations and water level data

We selected nine existing spots in Austria and Switzerland where water levels were measured by agencies or research groups at a nearby location (<21 km away; median: 0.2 km away) in the same stream (Figure 2; Table 1). The selected spots had at least one year of data by October 2019 and at least 45 or more contributions (ranging from 46 for spot A8, Rhine – Sevelen to 505 for Spot A2 Königseeache – Hallein). Furthermore, at twelve locations in Switzerland we installed signposts (*Figure 1d*) with reference images with the virtual staff gauge (Figure 2). On the signposts, we asked people to take a form, to record the WL-class, and to leave the form in the letterbox. We also asked the participants to record the date and time and whether they had participated in the CrowdWater project before. These stations are hereafter referred to as “pen-and-paper stations”. Two of the twelve pen-and-paper stations (P8 and P12) had a slightly different virtual staff gauge design than the one used in the app because they were installed early in the project and still had a prototype of the virtual staff gauge. Stations P1 and P4 are located at the same site as app spots A5 and A7, respectively, and have the same reference image. At P4 the staff gauge was set by us, but at A7 a citizen scientists created the spot. However, the largest difference between the app and the pen-and-paper stations was the number of citizen scientists who contributed to the observations for each station. The percentage of observations made by the citizen scientist who reported the most observations for a particular location varied between 74 and 100% for the app spots, and between <1 and 2% for the pen-and-paper stations, except for P4 (Limmat – Zürich; where 20% of the observations were submitted by the same person). Thus, for each app-spot, the majority of the observations were made by the same citizen scientist, whereas for the pen-and-paper stations almost every observation was made by a different person. The number of contributions for the selected app spots and pen-and-paper stations was similar (one observation every 1.2 to 11.6 days for the app spots (average: 5.3) vs an observation every 1.9 to 15.3 (average: 8.9) days for the pen–and-paper stations).

[Figure 2 here]

[Table 1 here]

## 2.3 Comparison of WL-class observations and measured water levels

The app automatically records the date and time of each observation. For the pen-and-paper stations, the observers were asked to record the local time on the paper form. This allowed us to compare the reported WL-class with the measured water level at the time of observation to assess the quality of the WL-class data. The water level was not always measured at exactly the same location as the WL-class observation but for the analysis only the timing of the variations in the water level needs to be the same. Each WL-class corresponds to a range of actual water levels but we do not know this range for each WL-class and location. Therefore, we created box-plots of the measured water levels at the time of the WL-class observation for each reported WL-class and compared them visually. In the perfect case, a higher reported WL-class should always correspond to a higher water level and each WL-class covers a fixed range of measured water levels (i.e., the ranges for the measured water levels for each WL-class do not overlap). We used a Kruskal-Wallis test to check whether there were significant differences between the water levels attributed to the individual WL-classes for each app spot and pen-and-paper station. Since this test showed that there were significant differences between the WL-classes for all locations, we used a Bonferroni posthoc test to compare the water levels of all class observations with each other. For this analysis, we checked which combinations of classes were significantly different but excluded WL-classes with fewer than five observations. The results were then grouped by the distance between the two tested classes.

We, furthermore, used the Kendall rank correlation coefficient, also called Kendall's $\tau$ (Kendall, 1990) to determine the correlation between all WL-class observations and the measured water levels. We chose the Kendall rank correlation instead of the Spearman rank correlation because it is considered to be more robust for data that includes many ties (Croux and Dehon, 2010), which is the case for WL-class observations. We used the Mann-Whitney U-test to compare the median Kendall's $\tau$ for the app spots and the pen-and-paper stations to assess whether the data quality for the two methods was similar or not.

## 2.4 Contribution times

Citizen science data can be biased in time (Courter et al., 2013). For example, citizen scientists may be more inclined to record observations during sunny periods when water levels are low. This affects the value of these data for hydrological model calibration (S. Etter et al., 2020). We, therefore, also analysed the date and time of the WL-class observations. In particular, we analysed how the observation frequencies varied throughout the year, during a week, and with the time of the day.

Furthermore, we compared the distribution of the measured water levels at the time of the crowd-based WL-class observations to the distribution of the water level data for the entire study period (Table 1) to see if the citizen scientist observations covered the range of high and low flow conditions. More specifically, we determined the percent of citizen observations that were above the 90$^{th}$ percentile and below the 10$^{th}$ percentile of the measured water levels.

# 3 Results

## 3.1 Data quality

For the app spots, a higher WL-class generally coincided with a higher measured water level, although different WL-classes were chosen for similar water levels so that the range of measured water levels for WL-classes overlapped (Figure *3*). Kendall's τ varied between 0.65 and 0.90 (with $p<0.01$), except for A9 (Urtene – Moosseedorf) for which Kendall's τ was 0.45 (Figure *3*). For the pen-and-paper stations, the correlation between the WL-class and measured water levels was poorer, with Kendall's τ values ranging between 0.05 and 0.57 (Figure 4). These Kendall's τ values were significantly lower than for the app spots ($p<0.01$; Figure 5). The overlap in the measured water levels for the observations for each WL-class was much larger for the pen-and-paper stations than the app-spots (Figure *3* and Figure 4). The Kruskal-Wallis and Bonferroni test result suggested that WL-classes that were further apart more often had significantly different median water levels (Table S*1*) and that this was more pronounced for the app data than the pen-and-paper data (cf. Figure *3*, Figure 4).

[Figure 3 here]

[Figure 4 here]

[Figure 5 here]

## 3.2 Contribution Times

On average, observations for the app spots were made throughout the daylight hours, although there was a tendency for more observations during the afternoons (Figure *6*). On average most contributions were made around 5pm but the differences between the spots were notable (Figure S2). Only some observations at the Limmat in Zurich (A7) were made outside the daylight hours (Figure S2). Observations were reported on both weekdays and weekends. However, there were significantly fewer contributions on Saturdays than the other days (on average 11% of all observations, whereas for all other days the average percentage varied between 14 and 16%; p=0.046; Figure *6*). Most WL-class observations were submitted during the warmer months of the year, i.e., between May and September (the average percentage of contributions per month varied between 10 and 11% for the May to September period, compared to 5 to 8% for the other months).

At the pen-and-paper stations, most observations were submitted in the early afternoon (Figure *6*). Furthermore, most contributions were received on Sundays (30% of all contributions; Figure *6* and Figure S3), followed by Saturdays (16%). Only 9 to 12% of the observations were submitted on the other days. Most observations were submitted in summer: more than 10 % of the observations were submitted for each month between May and August (except for July with 9.8 %), compared to 4 to 9% for the remaining months.

[Figure 6 here]

## 3.3 Range of WL-class observations

For the app spots, between 8 and 32% of the contributions (average: 16%) were submitted when the measured water level was above the 90$^{th}$ percentile; between 1 and 16% (average: 7%) of the observations were submitted when the measured water level was below the 10$^{th}$ percentile (Figure S4). At the pen-and-paper stations, observations were recorded less often during high water levels than for the app stations: between 0 and 20% of the observations (average: 11%) were made during times that the water level was above the 90$^{th}$ percentile. Between 0 and 23% of the observations (average: 9%) were submitted when the measured water level was below the 10$^{th}$ percentile (Figure S5).

# 4 Discussion

## 4.1 What is the quality of WL-class observations?

The WL-class observations made by citizen scientists with the CrowdWater app corresponded well with the measured water levels. Even though the results of such time series are not perfect and class boundaries are somewhat fuzzy, the estimated WL-classes from the app are well correlated with measured water levels. In some cases, a smaller staff gauge could have led to a higher resolution of the WL-class data in the spots A1, and A4 to A9. We assume that a similar number of covered classes as in A2 and A3 would not have decreased the data quality because A2 and A3 have the highest values for Kendall's τ (0.96 and 0.90). These results are thus encouraging for the CrowdWater project.

The observed WL-classes in the pen-and-paper stations did not correspond as well with the measured water levels as for the app spots. Even for the location at the Limmat in Zürich (P4) with 202 contributions made by 194 participants, the Kendall's τ is relatively low compared to the same spot in the app (A7) with only 73 contributions made by six participants (0.50 vs 0.71). The same is true for the Alp in Einsiedeln where we received 23 contributions made by 23 participants in P1 and 47 contributions by 8 participants in A5 (τ = 0.39 vs 0.69) and both stations had the exact same reference image. Furthermore, the differences between the individual classes were less often significant for the pen-and-paper stations than the app spots.

Strobl et al. (2019a) found, based on over 500 WL-class estimates using the same virtual staff gauge during surveys at ten rivers, that only 13% of the reported observations were more than one class off from the correct class (as determined by experts). Our results here are similar with respect to the very few outliers (Figure 3 and Figure 4). To some degree, errors in the use of the virtual staff gauge are to be expected because the water level is compared to the reference image by the citizen scientists. If the background on which the staff gauge is inserted is distorted, the comparison of reference structures to the WL-classes on the virtual staff gauge becomes more difficult (Seibert et al., 2019), especially when the water surface is not flat due to waves (e.g. A3) or if the riverbed is not clearly defined (e.g. in P2). Because previous choices of WL-classes at similar water levels were not easily visible in the app (i.e., only when one scrolls through the different observations) and not at all for the forms, the citizen scientists had few or no photos of similar conditions available to aid their decision on which WL-class to choose. Also, the virtual staff gauge approach is harder to understand than the approach of CrowdHydrology (Lowry et al., 2019) or the project in Kenya by Weeser et al. (2018) where water levels are read from physical staff gauges in, for instance, centimetres. This may contribute to poorer data quality for the novice contributors, and may explain the lower correlation with the measured water levels for the pen-and-paper contributions. Hence, the difference in data quality between the two approaches can be explained by the number of novice contributors: In the app, the data for each spot were mainly collected by a single dedicated person. If, the main contributor for an app spot has a constant bias (e.g., always estimating the water level too high), the time series would still be consistent. Since the virtual staff gauge approach largely builds on human perception, mistakes and less consistent results are more likely if there are many different contributors, especially if they are novice contributors. Furthermore, the different citizen scientists for the pen-and-paper stations likely all had a different bias. An alternative approach for the pen-and-paper stations could be that people at the pen-and-paper stations submit photographs of the actual situation to a server and then the WL-class can be estimated by a collective effort in, for instance, an online game. This is already possible for the app data (Strobl et al., 2019b).

The fuzzy separation of WL-classes based on measured water levels might also have other reasons than errors by the contributors. Even though the water level measurement stations can be considered well-maintained, errors in the stage measurements can not be entirely avoided. Horner et al. (2018) found errors in water level measurements in the order of 4 to 12% at six gauging stations in France. This indicates that the measured water levels, which are treated as error-free in this study, might contribute to the fuzziness of the class borders. Furthermore, the locations where the water levels were measured, were not at exactly the same location as the spots in the CrowdWater app (Table 1) and we did not correct for potential differences in the timing of water level variations. However, Kendall's τ was not correlated with the distance between the stations. The low Kendall's τ (0.45) at A9 (Urtene - Mooseedorf) might be explained by the small variations in water levels due to the presence of a regulated lake upstream from the station. The measured water levels were also influenced by a wastewater treatment plant, from which water entered the stream between the CrowdWater spot and the water level gauging station.

## 4.2 What are the characteristics of good spots for WL-class data observations?

Evaluation of the reference pictures and virtual staff gauges for the spots used in this study, allows us to draw some conclusions on the characteristics of spots that are likely to lead to good data. These are:

- The staff gauge size needs to be appropriate for the water level fluctuations, so that the variability in water levels spans several WL-classes (as an example we refer to the results for station A3 (Salzach – Salzburg) and A9 (Urtene – Moosseedorf; Figure *3*).
- Distinct features in the reference image are necessary to accurately identify changes in the water level. Vegetation can hinder the identification of these features and as this can block the view of reference features during certain seasons (see e.g. A6 in Figure 2 and Figure *3*; Seibert et al., 2019).
- For each spot, the data are contributed by one or few dedicated citizen scientists who feel responsible for the spot (see section 4.1).

## 4.3 When do citizen scientists contribute WL-class observations?

The WL-class observations were surprisingly uniformly distributed throughout the year, week and daylight hours. The contributions for the pen-and-paper stations were higher on weekends, especially on Sundays compared to the app stations, where the contributions were distributed more uniform throughout the week. The higher percentage of contributions on weekends for the pen-and-paper stations can be explained by the fact that these are opportunistic contributions when people saw the signposts (e.g., during a walk) and spontaneously decided to contribute. In two studies on citizen science reports of bird sightings such a “weekend-bias” was found to be stronger in Europe (Sparks et al., 2008) than the United States (Courter et al., 2013). Based on our data and own observations, Sunday seems to be the most likely day for people to be on such walks or hikes.

We assume that for the app spots used in this study, the contributors were more committed citizen scientists who included the submission of their observations as a part of a more regular routine (e.g., while going on a regular walk after work on the way to shops or walking the dog). One example is the small peak at 5 pm in the app stations, which might indicate that people contribute after work. However, this peak was influenced by the many contributions at A6 (Dünnern Balsthal) during this hour, for which almost 60% of the contributions were made between 5 and 6 pm. However, the contribution patterns varied notably between spots (Figure S2 and Figure S3), implying that it is hard to predict when dedicated citizen scientists will contribute.

The pen-and-paper stations received many responses when they were located at frequented paths, but people rarely contributed more than once. Potential reasons could be that people were only once at this location (i.e., during a one-time trip) or because they did not realise that multiple observations are helpful or because they missed feedback on their contribution. Feedback and visibility of participants contributions might lead to more sustained participation (Lowry et al., 2019). The app provides feedback to some extent by displaying all the contributions publicly. However, feedback on how the data are used and what individual contributions add to scientific research need to be communicated outside the app.

Loiselle et al. (2016) found that citizen scientists of the project FreshWaterWatch tended to make more repeated measurements if they get to choose the site for which they wanted to contribute data, compared to when stations were assigned to them. Furthermore, they also found that if many people contributed to the same stations, then the absolute number of contributions by a single contributor was smaller. This might to some extent be applicable to our study as well: People who see a signpost by chance and decide to contribute but feel less committed because there are potentially many others who could contribute than those who actively set up their own spot with the app and also can then check if there are other people contributing. Based on personal conversations with the main contributors to spots A2 and A4 and a motivation survey (Etter et al., in review), we assume that creating and maintaining own spots serves the needs for autonomy and competence. These are, in combination with the relatedness of one's own contributions to a broader topic, the needs to be fulfilled to foster self-determined and intrinsically motivated activism (Tiago et al., 2017) according to self-determination theory (Deci and Ryan, 2000). Frensley et al. (2017) argued that the motivation to participate in volunteering is increased if these three feelings are met. This would then lead to citizens who are motivated to observe high flows, and deliberately go out to do so. On the other hand, the pen-and-paper approach may lead to more interaction with the local population or a more diverse group of citizen scientists (Lowry et al., 2019).

## 4.4 Do the WL-class observations cover the entire range of water levels?

Our results show that the citizen scientists who use the app observed high and low flow conditions. In other words, the concern that the distribution of observed WL-classes might be biased to average or low flow conditions, or are otherwise fundamentally different from the long-term "true" distribution could not be confirmed. For the spot at the Alp in Einsiedeln (A5), 32% of the contributions were made at times when the water level was above the $90^{th}$ percentile. The main contributor for this spot stated in a personal conversation: "*The other day, I left the house again because it rained, to catch some high flows*." Thus, a citizen scientist who is particularly interested in high or low flows might provide data that contains information on extreme conditions as well.

For the pen-and-paper stations there were fewer contributions at high flows but rather more at low flows. This suggests that people who did not deliberately go outdoors to participate in the project are more likely to be outside and take time to submit their observations during periods with pleasant weather conditions. Therefore, to obtain observations over the entire range of water level conditions, it may be more beneficial to find dedicated citizen scientists than to catch the attention of many different citizen scientists.

# 5 Conclusions

The analysis shows that citizen scientists who use the CrowdWater app, were able to collect time series of WL-class data that are in good accordance with measured water levels (i.e., high correlation and few outliers). Observations for a spot submitted via the CrowdWater app by one or a few citizen scientists were of higher quality than the data from many different participants at the pen-and-paper stations. The uncertainties within the WL-classes could be due to mistakes of the citizen scientists but also due to the distance between the CrowdWater spots and the official gauging stations, as well as measurement errors.

The timing of the majority of the contributions for the app spots varied from site to site. The contributions with the app were made throughout the daylight hours but more frequently from May to September. Perhaps more importantly, the citizens submitted observations for all stream levels, including high water levels. The results are encouraging for citizen science in hydrology and demonstrate that with a smartphone app, dedicated volunteers can submit high quality water level class observations.

# 6 Acknowledgements

We thank all citizen scientists who contributed data to our app and pen-and-paper stations. Furthermore, we thank all the authorities and universities of Germany, Austria and Switzerland who allowed us to use their water level data: the State Departments of Hydrology of Niederösterriech and Salzburg, the Bavarian Hydrological Service, the Swiss Federal Office for the Environment (FOEN) and the Departments of Hydrometry for the cantons of Bern and Solothurn, as well as the Stream Biofilm and Ecosystem Research Laboratory of Tom Battin and Nicola Deluigi at the École Polytechnique Fédérale de Lausanne (EPFL). We, furthermore, thank Ronald Schmidt and the personnel of the

Wildnispark Zurich in Sihlwald, who helped to set up and maintain the pen-and-paper stations P8 and P12, and Nathalie Ceperley from the Université de Lausanne, who initiated the collaboration with the EPFL for the station in Vallon de Nant (P11). We also thank Hanspeter Hodel from the FOEN for maintaining the four pen-and-paper stations P2, P3, P6, and P9), and the Swiss National Park for the permission and the support with the station in the park (P5). The CrowdWater project is funded by the Swiss National Science Foundation (project no. 163008).

# 7 Data Availability Statement

The data that support the findings of this study are openly available in Zenodo (Etter et al., 2020; http://doi.org/10.5281/zenodo.3676351).

# 8 References

Balbo, A., Galimberti, G., 2016. Citizen hydrology in River Contracts for water management and people engagement at basin scale, in: COWM2016 - International Conference on Citizen Observatories for Water Management. Venice.

Courter, J.R., Johnson, R.J., Stuyck, C.M., Lang, B.A., Kaiser, E.W., 2013. Weekend bias in Citizen Science data reporting: implications for phenology studies. Int. J. Biometeorol. 57, 715–720. https://doi.org/10.1007/s00484-012-0598-7

Croux, C., Dehon, C., 2010. Influence functions of the Spearman and Kendall correlation measures. Stat. Methods Appt. 19, 497–515. https://doi.org/10.1007/s10260-010-0142-z

Davids, J.C., van de Giesen, N., Rutten, M., 2017. Continuity vs. the Crowd—Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations. Environ. Manage. 60, 12–29. https://doi.org/10.1007/s00267-017-0872-x

Deci, E.L., Ryan, R.M., 2000. The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior. Psychol. Inq. 11, 227–268.

https://doi.org/10.1207/S15327965PLI1104_01

Etter, S., Strobl, B., Seibert, J., Meerveld, H.J., 2020. Value of crowd-based water level class observations for hydrological model calibration. Water Resour. Res. https://doi.org/10.1029/2019WR026108

Etter, S., Strobl, B., Seibert, J., van Meerveld, I., 2018. Value of uncertain streamflow observations for hydrological modelling. Hydrol. Earth Syst. Sci. 22, 5243–5257. https://doi.org/https://doi.org/10.5194/hess-22-5243-2018

Etter, S., Strobl, B., van Meerveld, I. (H. J.., Seibert, J., 2020. Data and R-Scripts for “Quality and timing of crowd-based water level class observations.” https://doi.org/10.5281/zenodo.3676351

Etter, S., van Meerveld, H.J. (Ilja), Seibert, J., Strobl, B., Niebert, K., n.d. What motivates people to participate in environmental citizen science projects? Citiz. Sci. Theory Pract.

Frensley, T., Crall, A., Stern, M., Jordan, R., Gray, S., Prysby, M., Newman, G., Hmelo-Silver, C., Mellor, D., Huang, J., 2017. Bridging the Benefits of Online and Community Supported Citizen Science: A Case Study on Motivation and Retention with Conservation-Oriented Volunteers. Citiz. Sci. Theory Pract. 2, 4. https://doi.org/10.5334/cstp.84

Hannah, D.M., Demuth, S., van Lanen, H.A.J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., Tallaksen, L.M., 2011. Large-scale river flow archives: importance, current status and future needs. Hydrol. Process. 25, 1191–1200. https://doi.org/10.1002/hyp.7794

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H.K., Pierrefeu, G., 2018. Impact of Stage Measurement Errors on Streamflow Uncertainty. Water Resour. Res. 54, 1952–1976. https://doi.org/10.1002/2017WR022039

Kendall, M.G., 1990. Rank Correlation Methods., 5th ed, Journal of the Royal Statistical Society. Series A (General). Oxford University Press, London, UK & New York, NY.

Lanfranchi, V.., Wrigley, S.N.. N., Ireson, N.., Wehn, U.., Ciravegna, F.., 2014. Citizens' observatories for situation awareness in flooding. ISCRAM 2014 Conf. Proc. - 11th Int. Conf. Inf. Syst. Cris. Response Manag. 145–154.

Loiselle, S., Thornhill, I., Bailey, N., 2016. Citizen science: advantages of shallow versus deep participation. Front. Environ. Sci. 4. https://doi.org/10.3389/conf.FENVS.2016.01.00001

Lowry, C.S., Fienen, M.N., Hall, D.M., Stepenuck, K.F., Paul, J.D., 2019. Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology. Front. Earth Sci. 7, 1–10. https://doi.org/10.3389/feart.2019.00128

Mishra, A.K., Coulibaly, P., 2009. Developments in hydrometric network design: A review. Rev. Geophys. 47, RG2001. https://doi.org/10.1029/2007RG000243

Seibert, J., Strobl, B., Etter, S., Hummer, P., van Meerveld, H.J. (Ilja), 2019. Virtual Staff Gauges for Crowd-Based Stream Level Observations. Front. Earth Sci. 7. https://doi.org/10.3389/feart.2019.00070

Sivapalan, M., 2003. Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrol. Process. 17, 3163–3170. https://doi.org/10.1002/hyp.5155

Sparks, T.H., Huber, K., Tryjanowski, P., 2008. Something for the weekend? Examining the bias in avian phenological recording. Int. J. Biometeorol. 52, 505–510. https://doi.org/10.1007/s00484-008-0146-7

Strobl, B., Etter, S., van Meerveld, I., Seibert, J., 2019a. Accuracy of crowdsourced streamflow and stream level class estimates. Hydrol. Sci. J. 1–19. https://doi.org/10.1080/02626667.2019.1578966

Strobl, B., Etter, S., van Meerveld, I., Seibert, J., 2019b. The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. PLoS One 14, e0222579. https://doi.org/10.1371/journal.pone.0222579

Tiago, P., Gouveia, M.J., Capinha, C., Santos-Reis, M., Pereira, H.M., 2017. The influence of motivational factors on the frequency of participation in citizen science activities. Nat. Conserv. 18, 61–78. https://doi.org/10.3897/natureconservation.18.13429

Walker, D., Forsythe, N., Parkin, G., Gowing, J., 2016. Filling the observational void: Scientific value and quantitative validation of hydrometeorological data from a community-based monitoring programme. J. Hydrol. 538, 713–725. https://doi.org/10.1016/j.jhydrol.2016.04.062

Weeser, B., Stenfert Kroese, J., Jacobs, S.R., Njue, N., Kemboi, Z., Ran, A., Rufino, M.C., Breuer, L., 2018. Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. Sci. Total Environ. 631–632, 1590–1599. https://doi.org/10.1016/j.scitotenv.2018.03.130

# 9 Figures

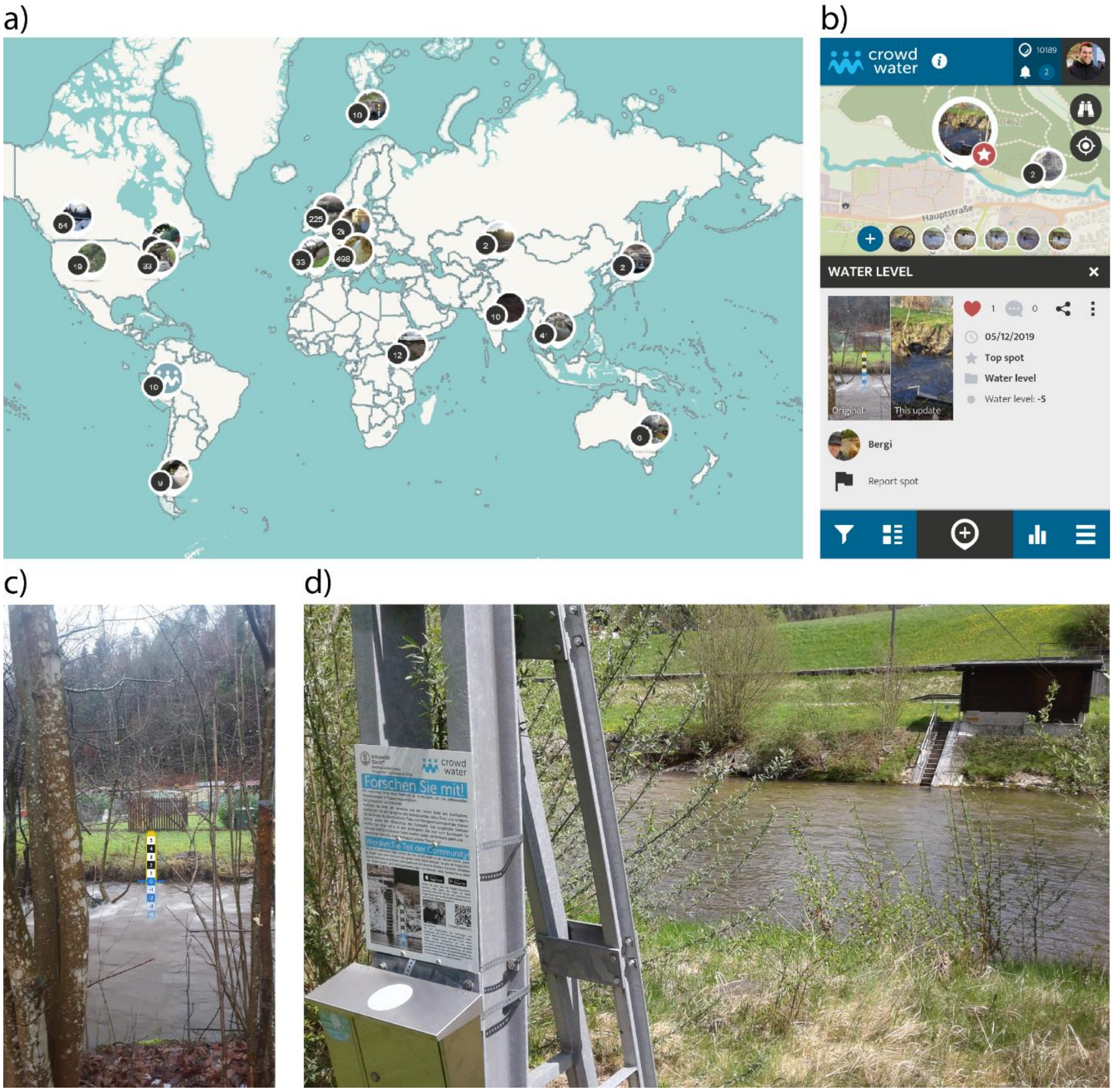



Figure 1 *Screenshot of the CrowdWater app showing the locations of existing spots on the map by 01.02.2019 (a), a screenshot showing the location of an existing spot, the reference picture with the virtual staff gauge and a photo of the current situation (b), a larger reference picture with the virtual staff-gauge (c), and a photo of the pen-and-paper station at the gauging station Kleine Emme – Werthenstein in Switzerland (P3) (d). In b, the image labelled "original" shows the reference picture with the virtual staff gauge (same image as in c) and the image labelled "This update" shows the new observation. Note also the reference image in the lower left of the signpost in d.*

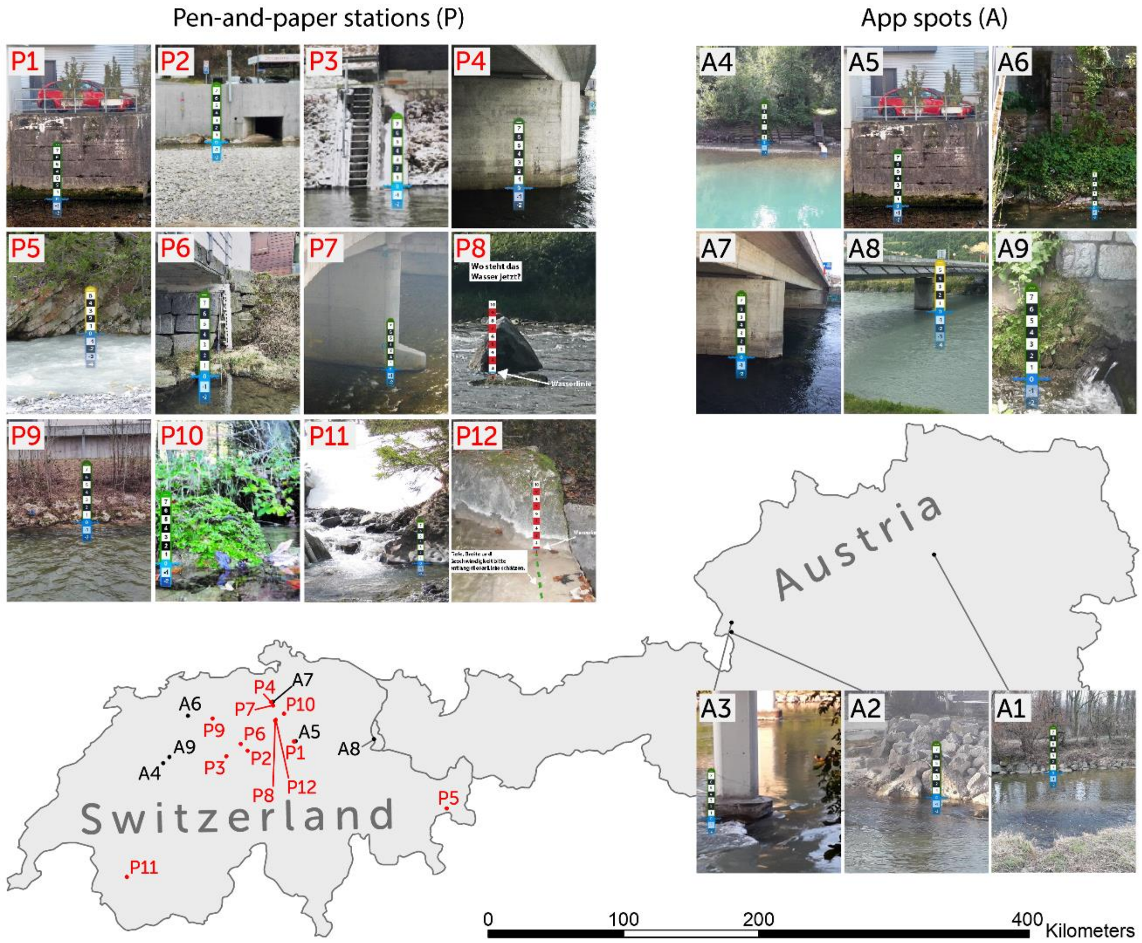



Figure 2 *Reference images with the virtual staff gauges for the app spots and pen-and-paper stations used in this study and their locations in Austria and Switzerland. Labels starting with "A" refer to app stations, labels starting with "P" refer to pen-and-paper stations. Note that the red and white staff gauges (in P8 and P12) are an early version of the staff gauge used in the app (Seibert et al., 2019).*

.

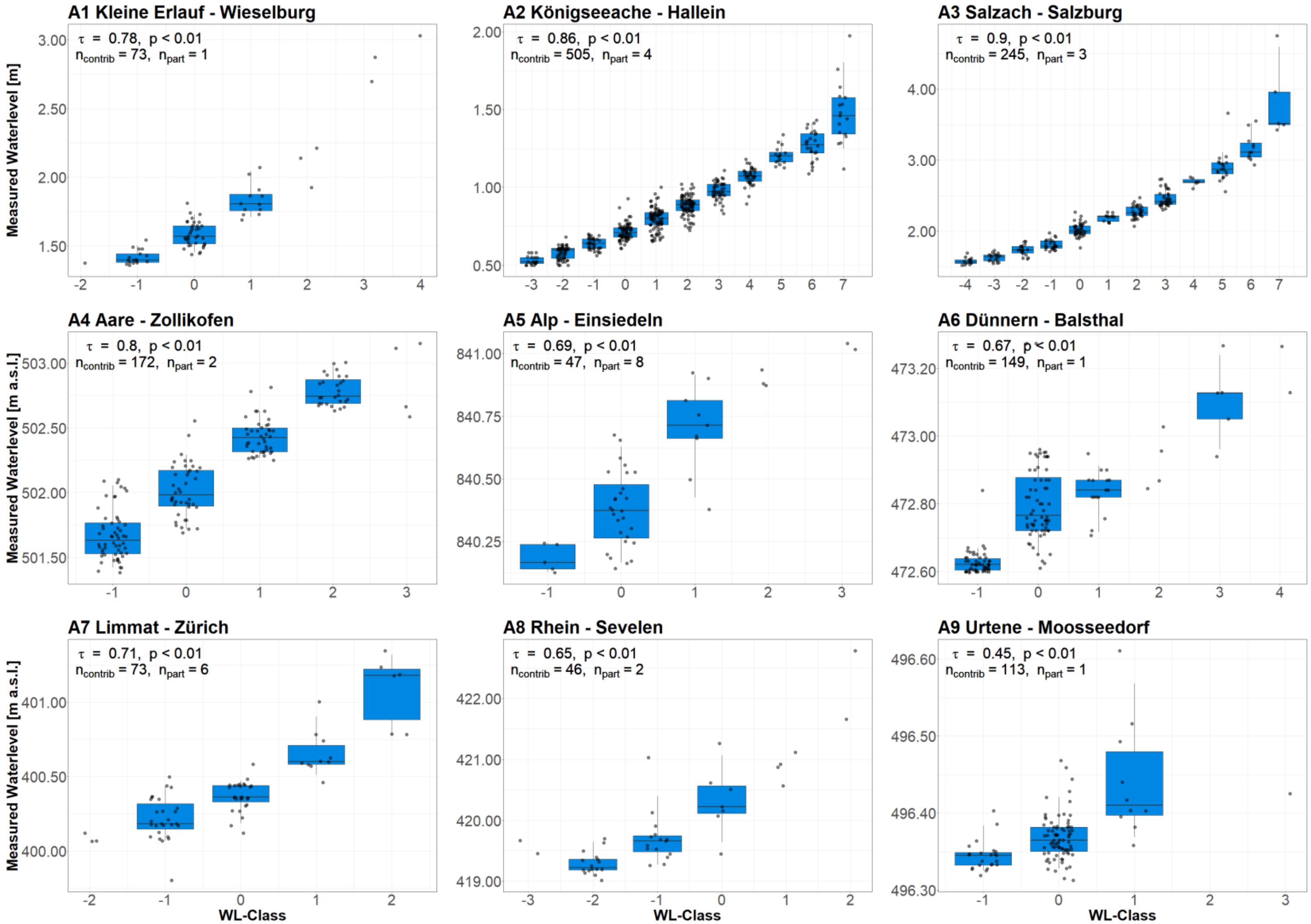



Figure 3 *Boxplots of the measured water levels at the time of a WL-class observation for each of the nine app spots. The box indicates the 25$^{th}$ to 75$^{th}$ percentile, the line the median, and the whiskers extend to the 5$^{th}$ and the 95$^{th}$ percentile. The dots (jittered) represent individual observations. τ is the correlation coefficient of Kendall's τ test, p the corresponding p-value. $n_{contrib}$ is the number of contributions (total number of dots), and $n_{part}$ the number of participants who contributed to the observations.*

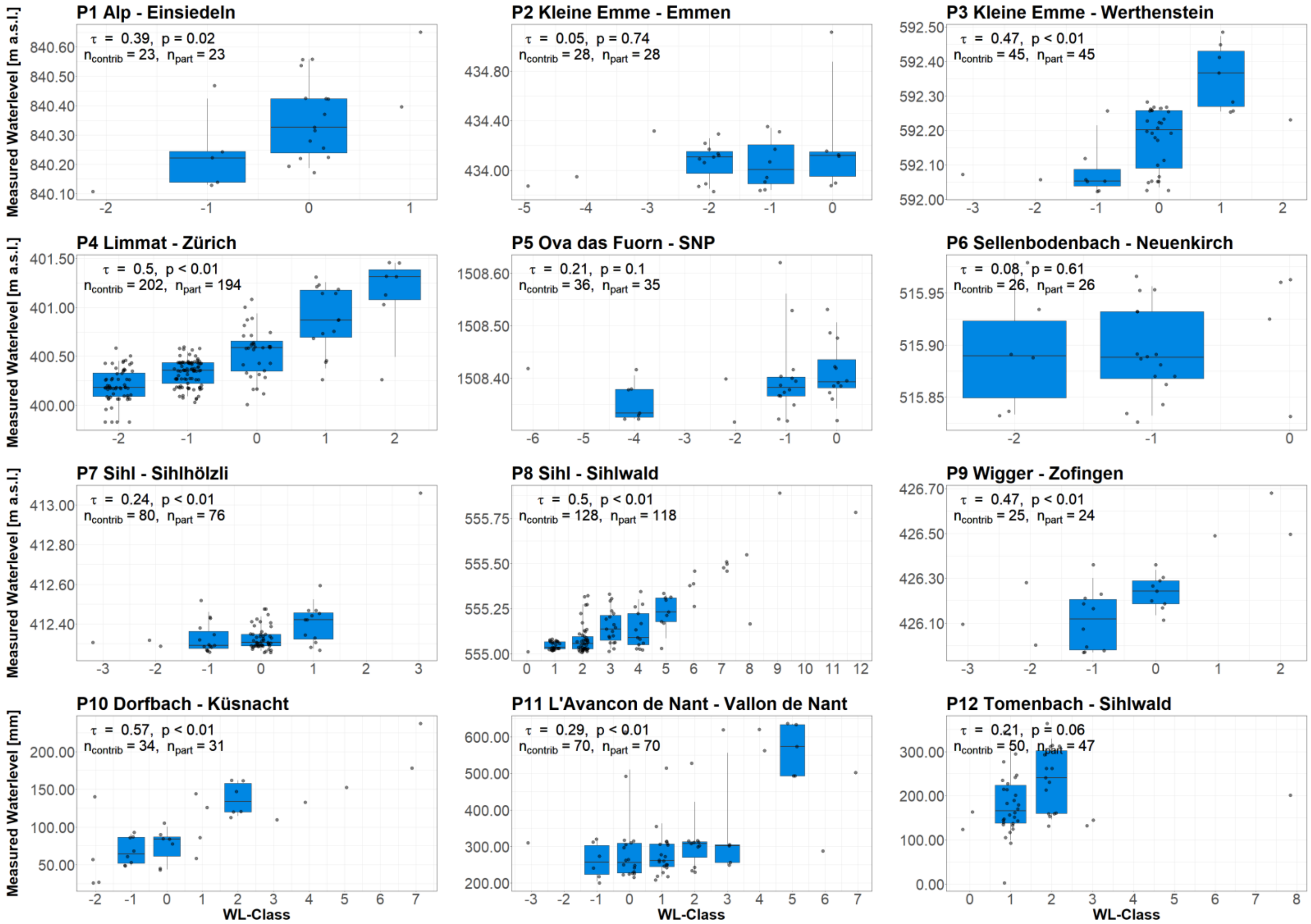



*Figure 4 Boxplots of the measured water level at the time of a WL-class observation for each of the twelve pen-and-paper stations. The box indicates the 25th to 75th percentile, the line the median, and the whiskers extend to the 5th and the 95th percentile. The dots (jittered) represent individual measurements. τ is the correlation coefficient of Kendall's τ test, p the corresponding p-value $n_{contrib}$ is the number of contributions (total number of dots), and $n_{part}$ the number of participants who contributed to the observations.*

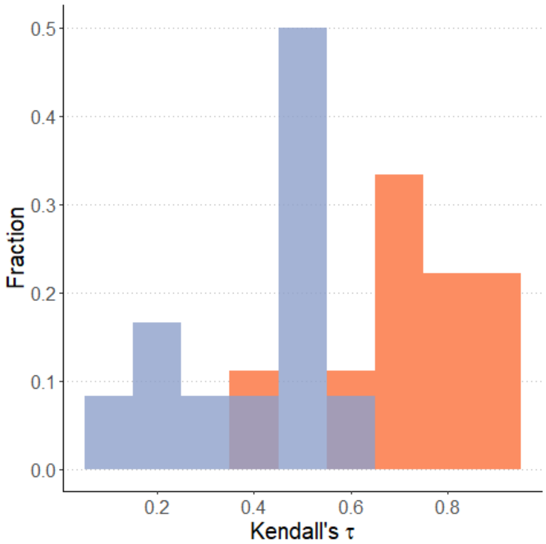



*Figure 5 Frequency distribution of the Kendall's τ for the relation between the WL-class observations and the measured water levels for the nine app spots (orange) and twelve pen-and-paper stations (purple) analysed in this study*

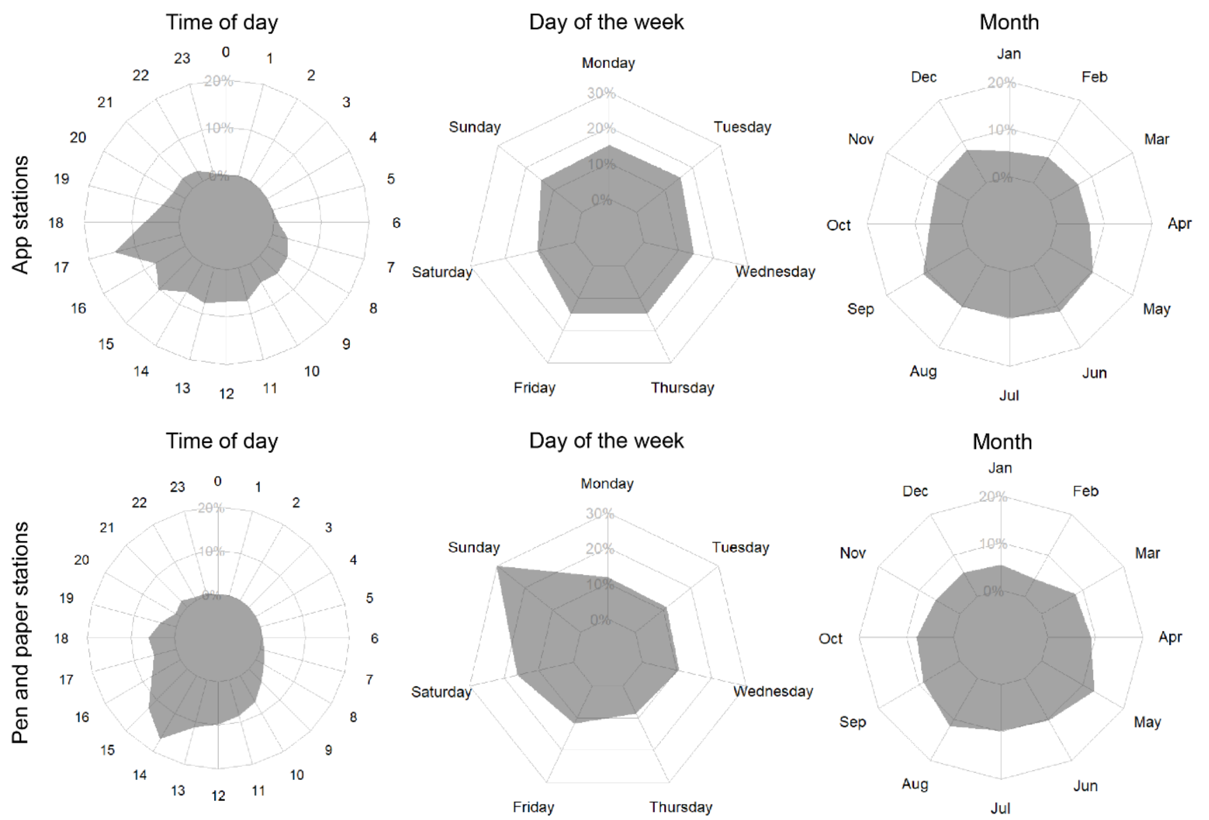



Figure 6 *Rose diagrams showing the average percentage of contributions for all nine app-spots (top) and pen-and-paper stations (bottom) analysed in this study for each time of the day (left), day of the week (middle), and month of the year (right). The results for each individual app spot can be found in supplemental material in Figure S2 and for the pen-and-paper stations in Figure S3.*

# Tables

Table 1 *Names and coordinates (decimal degrees N and E) of the app spots and pen-and-paper stations used in this study, the location of the water level measurements, the number of observations, the number of contributors, the correlation between the WL-class observations and the measured water levels (Kendall's ▯) and the corresponding p-values. The water level data were obtained from the state departments of hydrology of Niederösterriech (Amt der Niederösterreichischen Landesregierung – Abteilung für Hydrologie und Geoinformation; NOE) and Salzburg (German: Amt der Salzburger Landesregierung – Abteilung Wasser; ASL), the Bavarian Hydrological Service (Gewässerkundlicher Dienst Bayern; GKD), the Swiss Federal Office for the Environment (FOEN), the Departments of Hydrometry for two Swiss cantons, or our own measurements using Keller DCX-22 pressure sensors and water levels measured by the École Polytechnique Fédérale de Lausanne (EPFL) using TruTrack WT-HR 1000 water level loggers.*

| **Number** | **Station Name** | **Observation period** | **Coordinates WL measure-ments [N, E]** | **Source water level data** | **Coordinates WL-class observations [N, E]** | **Distance between WL and WL-class locations [km]** | **Number of obser-vations** | **Number of partici-pants** | **Kendall's τ** | **p-value** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | **App spots in Austria** | | | | | |
| A1 | Kleine Erlauf - Wieselburg | 30.03.2018 – 02.08.2019 | 48.1273, 15.1330 | NOE | 48.1255, 15.1292 | 0.3 | 73 | 1 | 0.78 | <0.01 |
| A2 | Königseeache - Hallein | 05.01.2018 – 10.09.2018 | 47.6458, 13.0303 | GKD | 47.7261, 13.0650 | 9.3 | 505 | 4 | 0.86 | <0.01 |
| A3 | Salzach - Salzburg | 26.08.2018 – 21.09.2019 | 47.7982, 13.0539 | ASL | 47.7896, 13.0686 | 1.5 | 245 | 3 | 0.90 | <0.01 |
| | | | | | **App spots in Switzerland** | | | | | |
| A4 | Aare - Zollikofen | 10.09.2017 – 30.04.2019 | 46.9333, 7.4480 | FOEN | 46.9904, 7.4508 | 6.4 | 172 | 2 | 0.80 | <0.01 |
| A5 | Alp-Einsiedeln | 29.11.2017 – 30.05.2019 | 47.1508, 8.7393 | FOEN | 47.1277, 8.7432 | 2.6 | 47 | 8 | 0.69 | <0.01 |
| A6 | Dünnern-Balsthal | 19.06.2018 – 22.06.2019 | 47.3022, 7.6975 | Canton of Solothurn | 47.3034, 7.6950 | 0.2 | 149 | 1 | 0.67 | <0.01 |
| A7 | Limmat-Zürich | 05.05.2017 – 17.02.2019 | 47.3908, 8.5257 | FOEN | 47.3919, 8.5233 | 0.2 | 73 | 6 | 0.71 | <0.01 |
| A8 | Rhein-Sevelen | 26.05.2018 – 11.06.2019 | 47.3067, 9.5710 | FOEN | 47.1301, 9.5114 | 20.2 | 46 | 2 | 0.65 | <0.01 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A9 | Urtene Moosseedorf | 21.06.2018 – 27.06.2019 | 47.0728, 7.5426 | Canton of Bern | 47.0301, 7.5116 | 5.3 | 113 | 1 | 0.45 | <0.01 |
| **Pen-and-paper stations (all in Switzerland)** | | | | | | | | | | |
| P1 | Alp - Einsiedeln | 10.03.2018 – 01.11.2018 | 47.1508, 8.7393 | FOEN | 47.1277, 8.7432 | 2.6 | 23 | 23 | 0.39 | 0.02 |
| P2 | Kleine Emme – Emmen | 30.04.2018 – 30.05.2019 | 47.0706, 8.2773 | FOEN | 47.0706, 8.2773 | 0.0 | 28 | 28 | 0.05 | 0.74 |
| P3 | Kleine Emme – Werthenstein | 28.04.2018 – 30.05.2019 | 47.0349, 8.0685 | FOEN | 47.0349, 8.0681 | 0.0 | 45 | 45 | 0.47 | <0.01 |
| P4 | Limmat – Zürich | 22.05.2017 – 15.06.2018 | 47.3906, 8.5254 | FOEN | 47.3918, 8.5234 | 0.2 | 202 | 194 | 0.50 | <0.01 |
| P5 | Ova da Fuorn – Swiss national Park | 12.08.2017 – 21.10.2017 | 46.6551, 10.1900 | FOEN | 46.6568, 10.1927 | 0.3 | 36 | 35 | 0.21 | 0.10 |
| P6 | Sellenbodenbach – Neuenkirch | 04.05.2018 – 24.05.2019 | 7.1128, 8.2102 | FOEN | 7.1128, 8.2102 | 0.0 | 26 | 26 | 0.08 | 0.61 |
| P7 | Sihl – Sihlhölzli | 11.05.2017 – 21.07.2018 | 47.3678, 8.5262 | FOEN | 47.3690, 8.5280 | 0.2 | 80 | 76 | 0.24 | 0.01 |
| P8 | Sihl – Sihlwald | 16.10.2016 – 12.05.2019 | 47.3678, 8.5262 | FOEN | 47.2714, 8.5566 | 11.0 | 128 | 118 | 0.50 | <0.01 |
| P9 | Wigger – Zofingen | 03.05.2018 – 21.05.2019 | 47.2836, 7.9350 | FOEN | 47.2836, 7.9354 | 0.0 | 25 | 24 | 0.47 | <0.01 |
| P10 | Dorfbach – Küsnacht | 30.11.2017 – 26.12.2018 | 47.3126, 8.6338 | pressure sensor | 47.3126, 8.6333 | 0.0 | 34 | 31 | 0.57 | <0.01 |
| P11 | L'Avancon de Nant – Vallon de Nant | 04.05.2018 – 26.06.2019 | 46.2315, 7.1019 | water level logger | 46.2316, 7.1022 | 0.0 | 70 | 70 | 0.29 | <0.01 |
| P12 | Tomenbach – Sihlwald | 28.01.2018 – 22.04.2019 | 47.2678, 8.5460 | pressure sensor | 47.2684, 8.5476 | 0.1 | 50 | 47 | 0.21 | 0.06 |

# 10 Supplemental Material

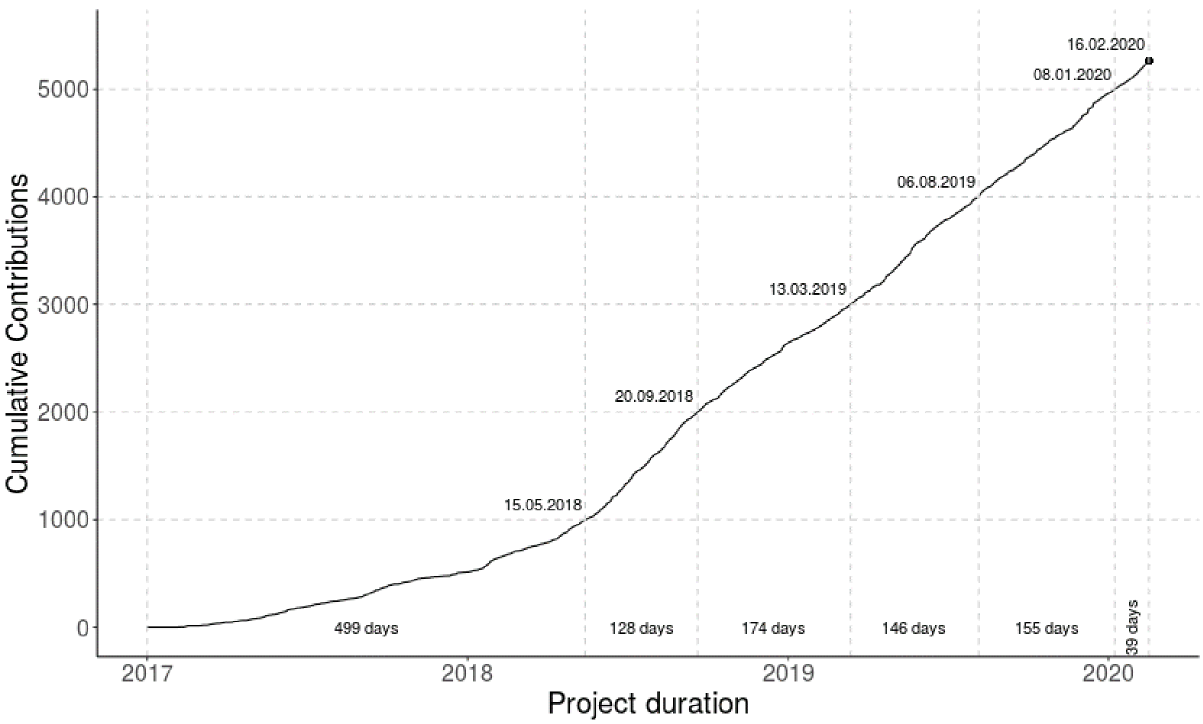



*Figure S1 Cumulative number of water level class observations submitted via the CrowdWater app. Figure obtained from crowdwater.ch/dashboard. Accessed: 16.02.2020.*

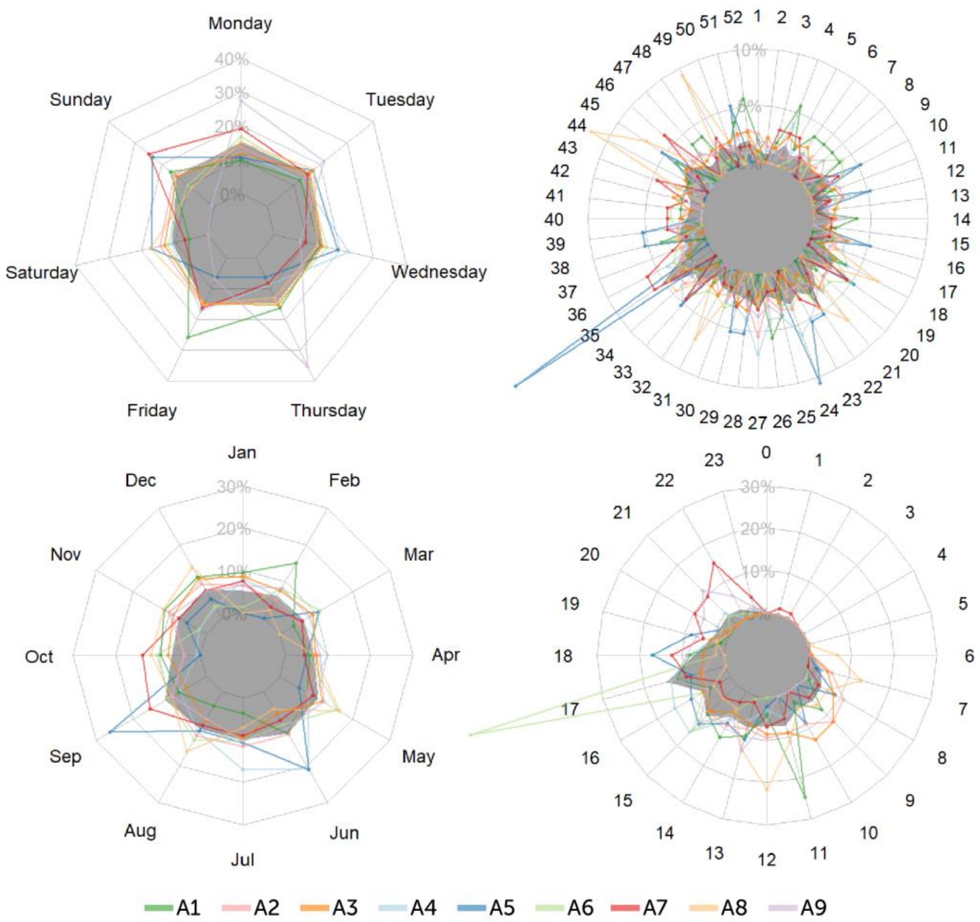



*Figure S2 Distribution of the time of all contributions for the individual app spots used in this study (lines) and the average for all spots (grey area, as reported in Figure 6): day of week (a), week of year (b), month of year (c) and hour of day (d).*

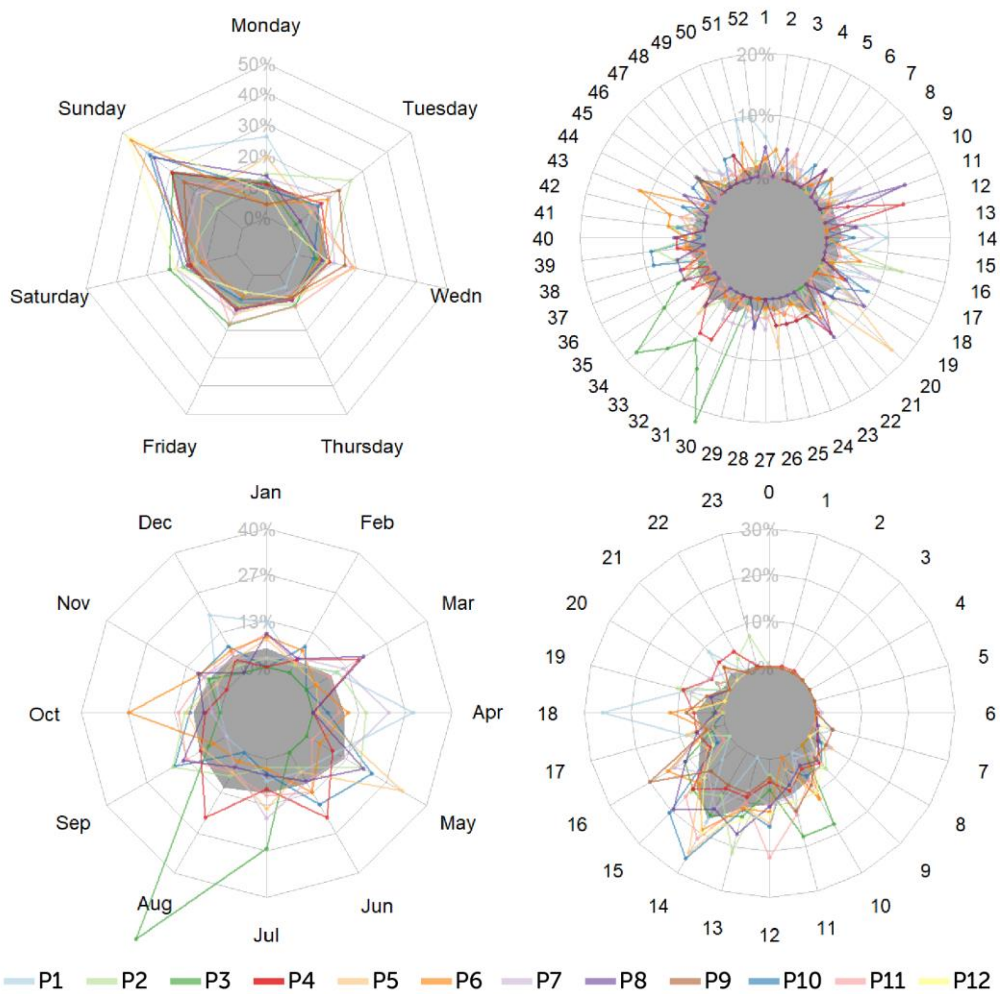



*Figure S3 Distribution of the time of all contributions for the individual pen-and paper stations used in this study (lines) and the average for all spots (grey area, as reported in* Figure 6*): day of week (a), week of year (b), month of year (c) and hour of day (d). Note that the weekly (b) and monthly (c) distributions are not plotted for stations for which less than 1 year of data were available.*

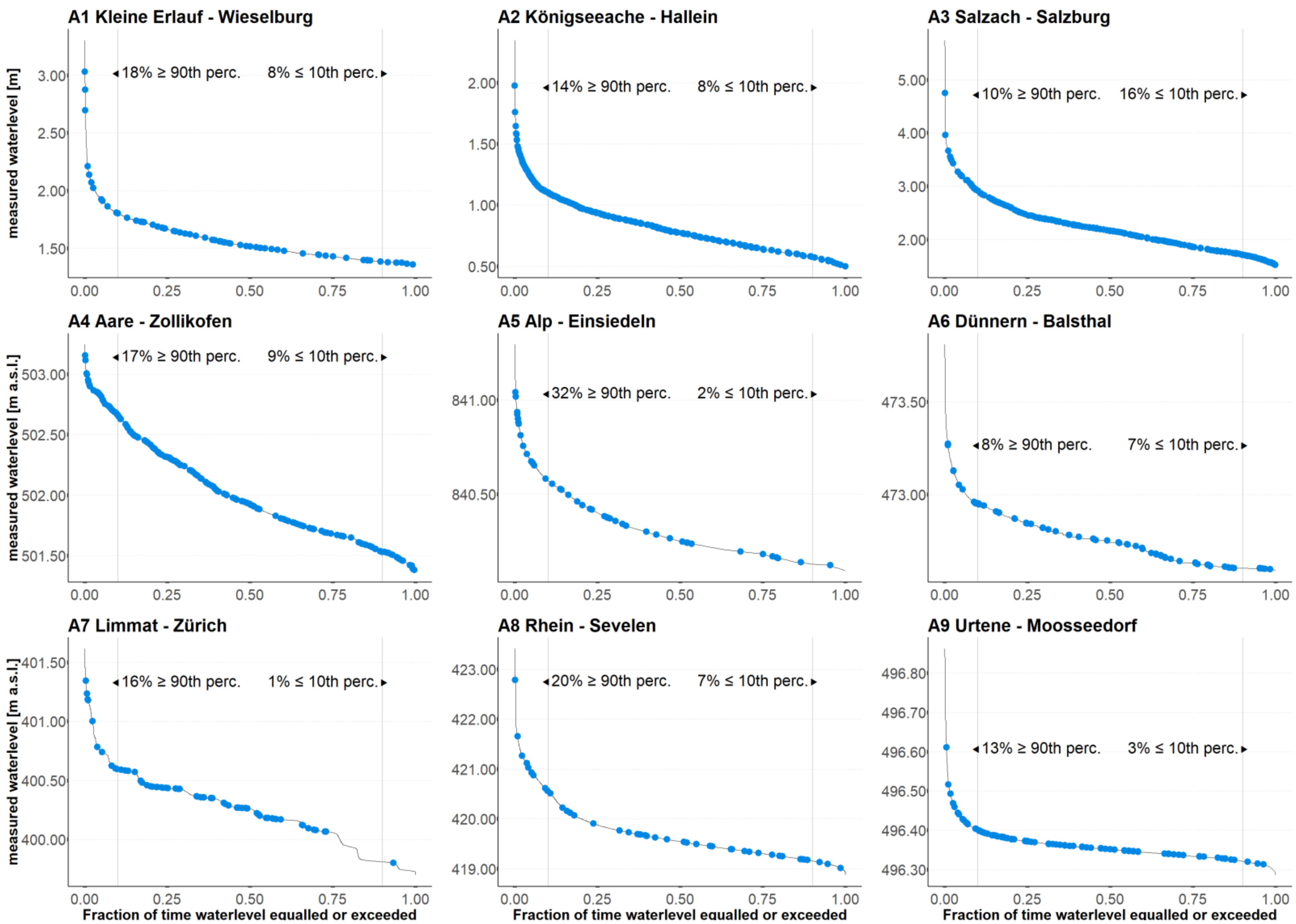



Figure S4 *The fraction of time that the water level was equal or exceeded (i.e. water level duration curve) at the official gauging stations (black lines) and the water level at the time of a WL-class observation submitted via the app (blue points). Both datasets cover the same period, i.e. the first and the last considered WL-class observations.*

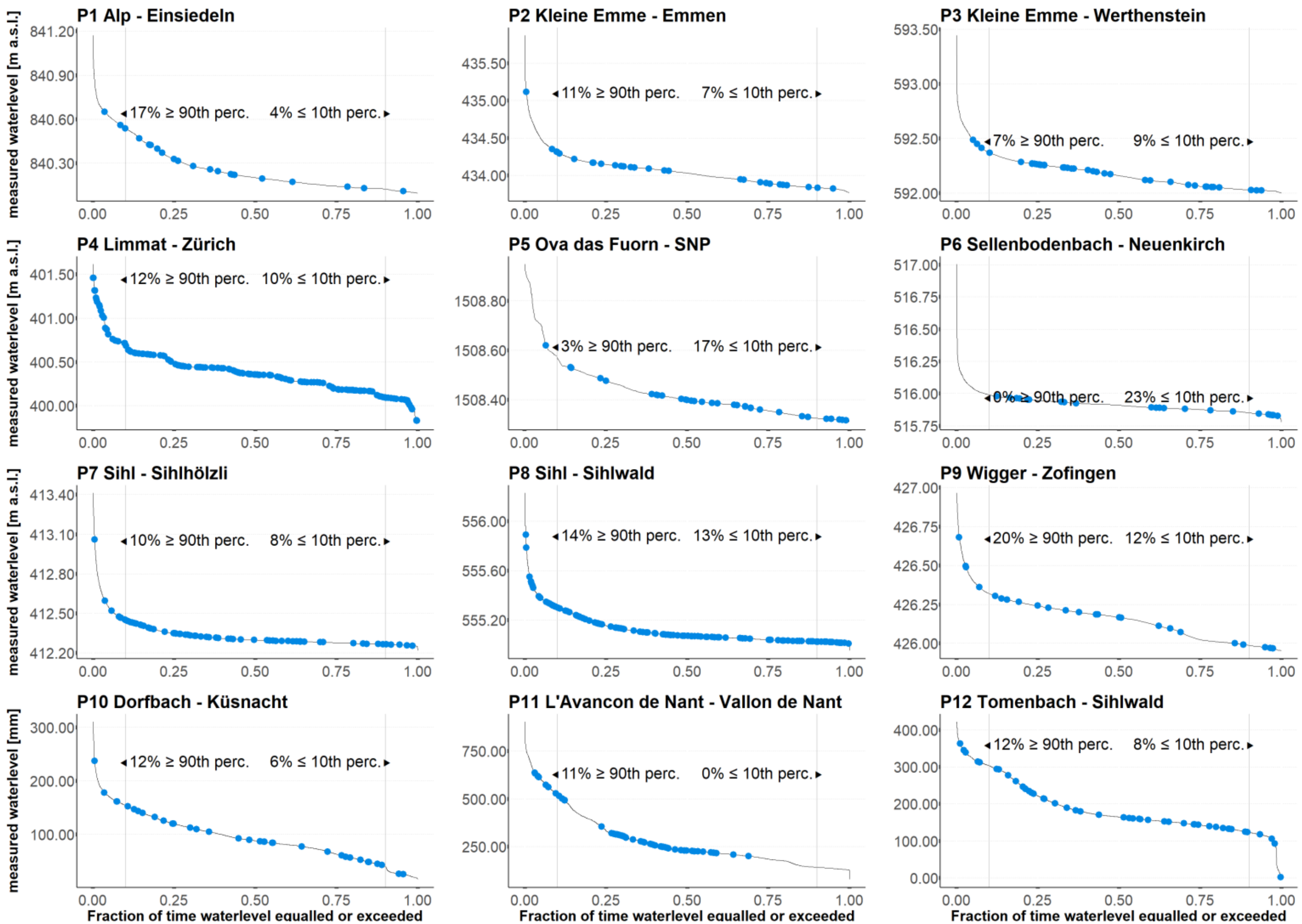



Figure S5 *The fraction of time that the water level was equal or exceeded (i.e. water level duration curve) at the official gauging stations (black lines) and the water level at the time of a WL-class observation submitted at a pen-and-paper station (blue points). Both datasets cover the same period, i.e. the first and the last considered WL-class observations.*

Table S1 *The fraction of significant class differences per app spot and pen-and-paper station based on the adjusted p-values from the Bonferroni test. The column names indicate the distance between classes (i.e. class 1 and 2 are 1 class apart, whereas class 1 and 3 are 2 classes apart). Note that the different spots and stations have different numbers of classes and therefore different distances could be covered in this analysis. Only WL-classes with five or more observations were included.*

| | **Distance between classes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Fraction of classes with significant differences with given distance** | | | | | | | | | | |
| | | **App spots** | | | | | | | | | | |
| A1 | Kleine Erlauf - Wieselburg | 1 | 1 | | | | | | | | | |
| A2 | Königseeache - Hallein | 0.1 | 0.56 | 0.75 | 0.86 | 1 | 1 | 1 | 1 | 1 | 1 | |
| A3 | Salzach - Salzburg | 0 | 0 | 0.56 | 0.88 | 0.86 | 0.83 | 1 | 1 | 1 | 1 | 1 |
| A4 | Aare - Zollikofen | 1 | 1 | 1 | | | | | | | | |
| A5 | Alp - Einsiedeln | 1 | 1 | | | | | | | | | |
| A6 | Dünnern - Balsthal | 0.50 | 0.5 | 0 | 1 | | | | | | | |
| A7 | Limmat - Zürich | 0.67 | 1 | 1 | | | | | | | | |
| A8 | Rhein - Sevelen | 0.50 | 1 | | | | | | | | | |
| A9 | Urtene - Moosseedorf | 1 | 1 | | | | | | | | | |
| | **Mean** | **0.64** | **0.78** | **0.66** | **0.91** | **0.93** | **0.92** | **1** | **1** | **1** | **1** | **1** |
| | **Median** | **0.67** | **1** | **0.75** | **0.88** | **0.93** | **0.92** | **1** | **1** | **1** | **1** | **1** |
| | | **Pen-and-paper stations** | | | | | | | | | | |
| P1 | Alp - Einsiedeln | 0 | | | | | | | | | | |
| P2 | Kleine Emme - Emmen | 0 | 0 | | | | | | | | | |
| P3 | Kleine Emme - Werthenstein | 0.50 | 1 | | | | | | | | | |
| P4 | Limmat - Zürich | 0.50 | 0.67 | 1 | 1 | | | | | | | |
| P5 | Ova da Fuorn - SNP | 0 | 0 | 0 | | | | | | | | |
| P6 | Sellenbodenbach - Neuenkirch | 0 | | | | | | | | | | |
| P7 | Sihl - Sihlhölzli | 0.50 | 1 | | | | | | | | | |
| P8 | Sihl - Sihlwald | 0.25 | 0.33 | 1 | 1 | | | | | | | |
| P9 | Wigger – Zofingen | 1 | | | | | | | | | | |
| P10 | Dorfbach - Küsnacht | 0 | 1 | 1 | | | | | | | | |
| P11 | L'Avancon de Nant – V. d. N. | 0 | 0 | 0 | 0.50 | 1 | 1 | | | | | |
| P12 | Tomenbach - Sihlwald | 1 | | | | | | | | | | |
| | **Mean** | **0.31** | **0.50** | **0.60** | **.83** | **1** | **1** | | | | | |
| | **Median** | **0.12** | **0.50** | **1** | **1** | **1** | **1** | | | | | |

# Paper V

Hydrol. Earth Syst. Sci., 22, 5243–5257, 2018
https://doi.org/10.5194/hess-22-5243-2018


Hydrology and Earth System Sciences
Open Access

# Value of uncertain streamflow observations for hydrological modelling

**Simon Etter**[1], **Barbara Strobl**[1], **Jan Seibert**[1,2], **and H. J. Ilja van Meerveld**[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, 75007 Uppsala, Sweden.

**Correspondence:** Simon Etter (simon.etter@geo.uzh.ch)

Received: 28 June 2018 – Discussion started: 11 July 2018
Revised: 20 September 2018 – Accepted: 24 September 2018 – Published: 15 October 2018

**Abstract.** Previous studies have shown that hydrological models can be parameterised using a limited number of streamflow measurements. Citizen science projects can collect such data for otherwise ungauged catchments but an important question is whether these observations are informative given that these streamflow estimates will be uncertain. We assess the value of inaccurate streamflow estimates for calibration of a simple bucket-type runoff model for six Swiss catchments. We pretended that only a few observations were available and that these were affected by different levels of inaccuracy. The level of inaccuracy was based on a log-normal error distribution that was fitted to streamflow estimates of 136 citizens for medium-sized streams. Two additional levels of inaccuracy, for which the standard deviation of the error distribution was divided by 2 and 4, were used as well. Based on these error distributions, random errors were added to the measured hourly streamflow data. New time series with different temporal resolutions were created from these synthetic streamflow time series. These included scenarios with one observation each week or month, as well as scenarios that are more realistic for crowdsourced data that generally have an irregular distribution of data points throughout the year, or focus on a particular season. The model was then calibrated for the six catchments using the synthetic time series for a dry, an average and a wet year. The performance of the calibrated models was evaluated based on the measured hourly streamflow time series. The results indicate that streamflow estimates from untrained citizens are not informative for model calibration. However, if the errors can be reduced, the estimates are informative and useful for model calibration. As expected, the model performance increased when the number of observations used for calibration increased. The model performance was also better when the observations were more evenly distributed throughout the year. This study indicates that uncertain streamflow estimates can be useful for model calibration but that the estimates by citizen scientists need to be improved by training or more advanced data filtering before they are useful for model calibration.

## 1 Introduction

The application of hydrological models usually requires several years of precipitation, temperature and streamflow data for calibration, but these data are only available for a limited number of catchments. Therefore, several studies have addressed the question: how many data points are needed to calibrate a model for a catchment? Yapo et al. (1996) and Vrugt et al. (2006), using stable parameters as a criterion for satisfying model performance, concluded that most of the information to calibrate a model is contained in 2–3 years of continuous streamflow data and that no more value is added when using more than 8 years of data. Perrin et al. (2007), using the Nash–Sutcliffe efficiency criterion (NSE), showed that streamflow data for 350 randomly sampled days out of a 39-year period were sufficient to obtain robust model parameter values for two bucket-type models, TOPMO, which is derived from TOPMODEL concepts (Michel et al., 2003), and GR4J (Perrin et al., 2003). Brath et al. (2004), using the volume error, relative peak error and time-to-peak error, concluded that at least 3 months of continuous data were required to obtain a reliable calibration. Other studies have shown that discontinuous streamflow data can be informative for constraining model parameters (Juston et al., 2009; Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015). Juston et al. (2009) used a multi-objective calibration that included groundwater data and concluded that the information content of a subset of 53 days of streamflow data was the same as for the 1065 days of data from which the subset was drawn. Seibert and Beven (2009), using the NSE criterion, found that model performance reached a plateau for 8–16 streamflow measurements collected throughout a 1-year period. They furthermore showed that the use of streamflow data for one event and the corresponding recession resulted in a similar calibration performance as the six highest measured streamflow values during a 2-month period.

These studies had different foci and used different model performance metrics, but nevertheless their results are encouraging for the calibration of hydrological models for ungauged basins based on a limited number of high-quality measurements. However, the question remains: how informative are low(er)-quality data? An alternative approach to high-quality streamflow measurements in ungauged catchments is to use citizen science. Citizen science has been proven to be a valuable tool to collect (Dickinson et al., 2010) or analyse (Koch and Stisen, 2017) various kinds of environmental data, including hydrological data (Buytaert et al., 2014). Citizen science approaches use simple methods to enable a large number of citizens to collect data and allow local communities to contribute data to support science and environmental management. Citizen science approaches can be particularly useful in light of the declining stream gauging networks (Ruhi et al., 2018; Shiklomanov et al., 2002) and to complement the existing monitoring networks. However, citizen science projects that collect streamflow or stream level data in flowing water bodies are still rare. Examples are the CrowdHydrology project (Lowry and Fienen, 2013), SmartPhones4Water in Nepal (Davids et al., 2018) and a project in Kenya (Weeser et al., 2018), which all ask citizens to read stream levels at staff gauges and to send these via an app or as a text message to a central database. Estimating streamflow is obviously more challenging than reading levels from a staff gauge but citizens can apply the stick or float method, where they measure the time it takes for a floating object (e.g. a small stick) to travel a given distance to estimate the flow velocity. Combined with estimates for the width and the average depth of the stream, this allows them to obtain a rough estimate of the streamflow. However, these streamflow estimates may be so inaccurate that they are not useful for model calibration. It is therefore necessary to not only evaluate the requirements of hydrological models in terms of the amount and temporal resolution of data, but also in terms of the achievable quality by the citizen scientists before starting a citizen science project.

The effects of rating curve uncertainty on model calibration (e.g. McMillan et al., 2010; Horner et al., 2018) and the value of sparse datasets (Davids et al., 2017) have been quantified in recent studies. However, the potential value of sparse datasets in combination with large uncertainties (such as those from crowdsourced streamflow estimates) has not been evaluated so far. Therefore, the aim of this study was to determine the effects of observation inaccuracies on the calibration of bucket-type hydrological models when only a limited number of observations are available. The specific objectives of this paper are to determine (i) whether the streamflow estimates from citizen scientists are informative for model calibration or if these errors need to be reduced (e.g. through training) to become useful and (ii) how the timing of the streamflow observations affects the calibration of a hydrological model. The latter is important for citizen science projects, as it provides guidance on whether it is useful to encourage citizens to contribute streamflow observations during a specific time of the year.

## 2 Methods

To assess the potential value of crowdsourced streamflow estimates for hydrological model calibration, the HBV (Hydrologiska Byråns Vattenbalansavdelning) model (Bergström, 1976) was calibrated against streamflow time series for six Swiss catchments, as well as for different subsets of the data that represent citizen science data in terms of errors and temporal resolution. Similar to the approach used in several recent studies (Ewen et al., 2008; Finger et al., 2015; Fitzner et al., 2013; Haberlandt and Sester, 2010; Seibert and Beven, 2009), we pretended that only a small subset of the data were available for model calibration. In addition, various degrees of inaccuracy were assumed. The value of these data for model calibration was then evaluated by comparing the model performance for these subsets of data to the performance of the model calibrated with the complete measured streamflow time series.

### 2.1 HBV model

The HBV model was originally developed at the Hydrologiska Byråns Vattenbalansavdelning unit at the Swedish Meteorological and Hydrological Institute (SMHI) by Bergström (1976). The HBV model is a bucket-type model that represents snow, soil, groundwater and stream routing processes in separate routines. In this study, we used the version HBV-light (Seibert and Vis, 2012).

### 2.2 Catchments

The HBV-light model was set up for six 24–186 $km^2$ catchments in Switzerland (Table 1 and Fig. 1). The catchments were selected based on the following criteria: (i) there is little anthropogenic influence, (ii) they are gauged at a single location, (iii) they have reliable streamflow data during high flow and low flow conditions (i.e. no complete freezing dur-

**Table 1.** Characteristics of the six Swiss catchments used in this study. For the location of the study catchments, see Fig. 1. Long-term averages are for the period 1974–2014, except for Verzasca for which the long-term average is for the 1990–2014 period. Regime types are classified according to Aschwanden and Weingartner (1985).

| Catchment | | Murg | Guerbe | Allenbach | Riale di Calneggia | Mentue | Verzasca |
|---|---|---|---|---|---|---|---|
| Gauging station (FOEN station number) | | Waengi (2126) | Belp Mülimatt (2159) | Adelboden (2232) | Cavergno, Pontit (2356) | Yvonand La Mauguettaz (2369) | Lavertezzo, Campiòi (2605) |
| Area (km$^2$) | | 79 | 117 | 29 | 24 | 105 | 186 |
| Elevation (m a.s.l.) | Min | 465 | 522 | 1297 | 885 | 445 | 490 |
| | Max | 1035 | 2176 | 2762 | 2921 | 927 | 2864 |
| Regime type | | Pluvial-inférieur | Pluvial-superieur | Nival-alpin | Nival-méridional | Pluvial-jurassien | Nivo-pluvial-méridional |
| Min–max Pardé coefficients | Dry year | 0.29–1.61 | 0.44–1.93 | 0.40–2.48 | 0.13–3.22 | 0.22–2.37 | 0.16–2.92 |
| | Average year | 0.58–2.16 | 0.61–1.65 | 0.39–2.44 | 0.09–2.84 | 0.23–2.66 | 0.23–3.17 |
| | Wet year | 0.34–1.69 | 0.42–2.14 | 0.32–2.12 | 0.10–3.48 | 0.35–2.39 | 0.26–2.64 |
| | Long-term | 0.68–1.34 | 0.77–1.39 | 0.35–2.70 | 0.14–2.70 | 0.46–1.57 | 0.23–2.22 |
| Annual runoff : rainfall ratio | Dry year | 0.72 | 0.37 | 0.86 | 1.30[1] | 0.41 | 0.98 |
| | Average year | 0.55 | 0.48 | 1.73[1] | 1.38[1] | 0.52 | 0.66 |
| | Wet year | 0.56 | 0.54 | 0.78 | 0.98 | 0.50 | 1.32[1] |
| | Long-term | 0.56 | 0.57 | 0.94 | 1.06[1] | 0.38 | 0.9 |
| Long-term mean annual streamflow (m$^3$ s$^{-1}$) | | 1.84 | 2.75 | 1.23 | 1.43 | 1.64 | 10.76 |
| Weather stations | | Aadorf-Taenikon, Hörnli | Plaffeien, Bern-Zollikofen | Adelboden | Robiei | Mathod, Pully | Acquarossa, Cimetta, Magadino, Piotta |

[1] In Verzasca, Allenbach and Riale die Calneggia there are some streamflow : rainfall ratios > 1 because the weather stations are located outside the catchment and precipitation is highly variable in alpine terrain.

ing winter and a cross section that allows accurate streamflow measurement at low flows) and (iv) there are no glaciers. The six selected catchments (Table 1) represent different streamflow regime types (Aschwanden and Weingartner, 1985). The snow-dominated highest elevation catchments (Allenbach and Riale di Calneggia) have the largest seasonality in streamflow, i.e. the biggest differences between the long-term maximum and minimum Pardé coefficients, followed by the rain- and snow-dominated Verzasca catchment. The rain-dominated catchments (Murg, Guerbe and Mentue) have the lowest seasonal variability in streamflow (Table 1). The mean elevation of the catchments varies from 652 to 2003 m a.s.l. (Table 1). The elevation range of each individual catchment was divided into 100 m elevation bands for the simulations.

### 2.3 Measured data

Hourly runoff time series (based on 10 min measurements) for the six study catchments were obtained from the Federal Office for the Environment (FOEN; see Table 1 for the gauging station numbers). The average hourly areal precipitation amounts were extracted for each study catchment from the gridded CombiPrecip dataset from MeteoSwiss (Sideris et al., 2014). This dataset combines gauge and radar precipitation measurements at an hourly timescale and 1 km$^2$ spatial resolution and is available for the time period since 2005.

We used hourly temperature data from the automatic monitoring network of MeteoSwiss (see Table 1 for the stations) and applied a gradient of $-6\,°C$ per 1000 m to adjust the temperature of each weather station to the mean elevation of the catchment. Within the HBV model, the temperature was then adjusted for the different elevation bands using a calibrated lapse rate.

As recommended by Oudin et al. (2005), potential evapotranspiration was calculated using the temperature-based potential evapotranspiration model of McGuinness and Bordne (1972) using the day of the year, the latitude and the temperature. This rather simplistic approach was considered

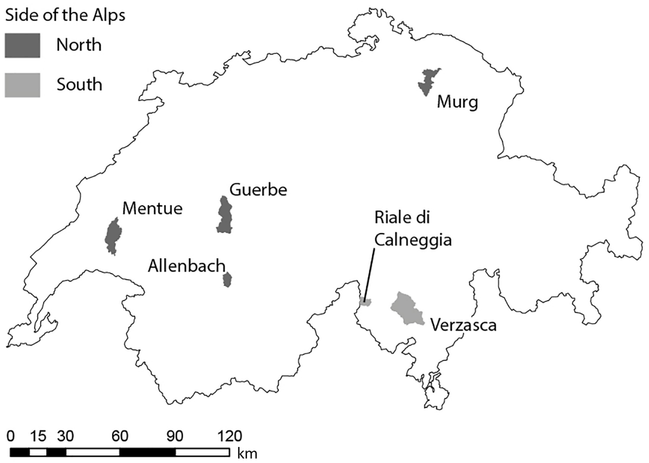



**Figure 1.** Location of the six study catchments in Switzerland. Shading indicates whether the catchment is located on the north or south side of the Alps. See Table 1 for the characteristics of the study catchments.

sufficient because this study focused on differences in model performance relative to a benchmark calibration.

### 2.4 Selection of years for model calibration and validation

The model was calibrated for an average, a dry and a wet year to investigate the influence of wetness conditions and the amount of streamflow on the calibration results. The years were selected based on the total streamflow during summer (July–September). The driest and the wettest years of the period 2006–2014 were selected based on the smallest and largest sum of streamflow during the summer. The average streamflow years were selected based on the proximity to the mean summer streamflow for all the years 1974–2014 (1990–2014 for Verzasca). For each catchment the years that were the 2nd-closest to the mean summer streamflow for all years, as well as the years with the second lowest and second highest streamflow sum were chosen for model calibration (see Table 2). We did this separately for each catchment because for each catchment a different year was dry, average or wet. For the validation, we chose the year closest to the mean summer streamflow and the years with the lowest and the highest total summer streamflow (see Table 2). We used each of the parameter sets obtained from calibration for the dry, average or wet years to validate the model for each of the three validation years, resulting in nine validation combinations for each catchment (and each dataset, as described below).

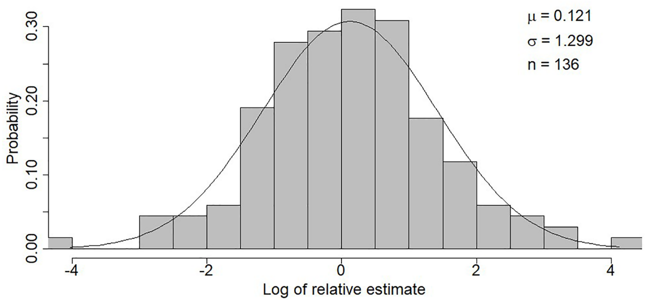



**Figure 2.** Fit of the normal distribution to the frequency distribution of the log-transformed relative streamflow estimates (ratio of the estimated streamflow and the measured streamflow).

### 2.5 Transformation of datasets to resemble citizen science data quality

#### 2.5.1 Errors in crowdsourced streamflow observations

Strobl et al. (2018) asked 517 participants to estimate streamflow based on the stick method at 10 streams in Switzerland. Here we use the estimates for the medium-sized streams Töss, Sihl and Schanzengraben in the Canton of Zurich and the Magliasina in Ticino ($n = 136$), which had a similar streamflow range at the time of the estimations (2.6–28 $m^3\,s^{-1}$) as the mean annual streamflow of the six streams used for this study (1.2–10.8 $m^3\,s^{-1}$). We calculated the streamflow from the estimated width, depth and flow velocities using a factor of 0.8 to adjust the surface flow velocity to the average velocity (Harrelson et al., 1994). The resulting streamflow estimates were normalised by dividing them by the measured streamflow. We then combined the normalised estimates of all four rivers and log-transformed the relative estimates. A normal distribution with a mean of 0.12 and a standard deviation of 1.30 fits the distribution of the log-transformed relative estimates well (standard error of the mean: 0.11, standard error of the standard deviation: 0.08; Fig. 2).

To create synthetic datasets with data quality characteristics that represent the observed crowdsourced streamflow estimates, we assumed that the errors in the streamflow estimates are uncorrelated (as they are likely provided by different people). For each time step, we randomly selected a relative error value from the log-normal distribution of the relative estimates (Fig. 2) and multiplied the measured streamflow with this relative error. To simulate the effect of training and to obtain time series with different data quality, two additional streamflow time series were created using a standard deviation divided by 2 (standard deviation of 0.65) and by 4 (standard deviation of 0.33). This reduces the spread in the data (but does not change the small systematic overestimation of the streamflow), so large outliers are still possible, but are less likely. To summarise, we tested the following four cases.

**Table 2.** The calibration years (second most extreme and second closest to average years) and validation years (most extreme and closest to average years) for each catchment. The numbers in parentheses are the ranks over the period 1974–2014 (or 1990–2014 for Verzasca).

| Year character | Murg | Guerbe | Allenbach | Riale di Calneggia | Mentue | Verzasca |
|---|---|---|---|---|---|---|
| Calibration | | | | | | |
| Wet | 2007 (3) | 2007 (2) | 2007 (4) | 2009 (11) | 2014 (7) | 2011 (4) |
| Dry | 2013 (8) | 2011 (8) | 2009 (11) | 2012 (8) | 2010 (4) | 2013 (5) |
| Average | 2008 (6) | 2008 (17) | 2013 (7) | 2013 (2) | 2006 (6) | 2007 (7) |
| Validation | | | | | | |
| Wet | 2014 (1) | 2014 (1) | 2014 (1) | 2008 (9) | 2007 (1) | 2008 (1) |
| Dry | 2009 (7) | 2013 (5) | 2012 (9) | 2006 (5) | 2009 (3) | 2010 (4) |
| Average | 2011 (4) | 2006 (13) | 2011 (6) | 2011 (1) | 2013 (2) | 2006 (4) |

– *No error*: the data measured by the FOEN, assumed to be (almost) error-free, the benchmark in terms of quality.

– *Small error*: random errors according to the log-normal distribution of the snapshot campaigns with the standard deviation divided by 4.

– *Medium error*: random errors according to the log-normal distribution of the surveys with the standard deviation divided by 2.

– *Large error*: typical errors of citizen scientists, i.e. random errors according to the log-normal distribution of errors from the surveys.

#### 2.5.2 Filtering of extreme outliers

Usually some form of quality control is used before citizen science data are analysed. Here, we used a very simple check to remove unrealistic outliers from the synthetic datasets. This check was based on the likely minimum and maximum streamflow for a given catchment area. We defined an upper limit of possible streamflow values as a function of the catchment area using the dataset of maximum streamflow from 1500 Swiss catchments provided by Scherrer AG, Hydrologie und Hochwasserschutz (2017). To account for the different precipitation intensities north and south of the Alps, different curves were created for the catchments on each side of the Alps. All streamflow observations, i.e. modified streamflow measurements, above the maximum observed streamflow for a particular catchment size including a 20 % buffer (Fig. S1), were replaced by the value of the maximum streamflow for a catchment of that size. This affected less than 0.5 % of all data points. A similar procedure was used for low flows based on a dataset of the FOEN with the lowest recorded mean streamflows over 7 days but this resulted in no replacements.

**Table 3.** Weights assigned to specific seasons, days and times of the day for the random selection of data points for Crowd52 and Crowd12. The weights for each hour were multiplied and normalised. We then used them as probabilities for the individual hours. For times without daylight the probability was set to zero.

| Variable | | Weight |
|---|---|---|
| Season | | |
| December–February | | 2 |
| March–May/September–November | | 6 |
| June–August | | 10 |
| Day | | |
| Saturdays–Sundays | | 3 |
| Monday–Friday | | 1 |
| Time | | |
| Times when people have breaks | 06:00–08:00, 12:00–13:00, 17:00–21:00 | 3 |
| Times with daylight in winter (December–February) | 08:00–16:00 | 1 |
| Times with daylight in spring/fall (March–May/September–November): | 07:00–19:00 | 1 |
| Times with daylight in summer (June–August) | 06:00–21:00 | 1 |
| Other times (depending on season) | | 0 |

#### 2.5.3 Temporal resolution of the observations

Data entries from citizen scientists are not as regular as data from sensors with a fixed temporal resolution. Therefore, we decided to test eight scenarios with a different temporal resolution and distribution of the data throughout the year to simulate different patterns in citizen contributions.

– *Hourly*: one data point per hour ($8760 \leq n \leq 8784$, depending on the year).

- *Weekly*: one data point per week, every Saturday, randomly between 06:00 and 20:00 ($52 \leq n \leq 53$).
- *Monthly*: one data point per month on the 15th of the month, randomly between 06:00 and 20:00 ($n = 12$).
- *IntenseSummer*: one data point every other day from July until September, randomly between 06:00 and 20:00 ($\sim 15$ observations per month, $n = 46$).
- *WeekendSummer*: one data point each Saturday and each Sunday between May and October, randomly between 06:00 and 20:00 ($52 \leq n \leq 54$).
- *WeekendSpring*: one data point on each Saturday and each Sunday between March and August, randomly between 06:00 and 20:00 ($52 \leq n \leq 54$).
- *Crowd52*: 52 random data points during daylight (in order to be comparable to the Weekly, IntenseSummer, WeekendSummer and WeekendSpring time series).
- *Crowd12*: 12 random data points during daylight (comparable to the Monthly data).

Except for the hourly data, these scenarios were based on our own experiences within the CrowdWater project (https://www.crowdwater.ch, last access: 3 October 2018) and information from the CrowdHydrology project (Lowry and Fienen, 2013). The hourly dataset was included to test the effect of errors when the temporal resolution of the data is optimal (i.e. by comparing simulations for the models calibrated with the hourly FOEN data and those calibrated with hourly data with errors). In the two scenarios Crowd52 and Crowd12, with random intervals between data points, we assigned higher probabilities for periods when people are more likely to be outdoors (i.e. higher probabilities for summer than winter, higher probabilities for weekends than weekdays, higher probabilities outside office hours; Table 3). Times without daylight (dependent on the season) were always excluded. We used the same selection of days, including the same times of the day for each of the four different error groups, years and catchments to allow comparison of the different model results.

## 2.6 Model calibration

For each of the 1728 cases (6 catchments, 3 calibration years, 4 error groups, 8 temporal resolutions), the HBV model was calibrated by optimising the overall consistency performance $P_{OA}$ (Finger et al., 2011) using a genetic optimisation algorithm (Seibert, 2000). The overall consistency performance $P_{OA}$ is the mean of four objective functions with an optimum value of 1: (i) NSE, (ii) the NSE for the logarithm of streamflow, (iii) the volume error and (iv) the mean absolute relative error (MARE). The parameters were calibrated within their typical ranges (see Table S1 in the Supplement.). To consider parameter uncertainty, the calibration was performed 100 times, which resulted in 100 parameter sets for each case. For each case, the preceding year was used for the warm-up period. For the Crowd52 and Crowd12 time series, we used 100 different random selections of times, whereas for the regularly spaced time series the same times were used for each case.

## 2.7 Model validation and analysis of the model results

The 100 parameters from the calibration for each case were used to run the model for the validation years (Table 2). For each case (i.e. each catchment, year, error magnitude and temporal resolution), we determined the median validation $P_{OA}$ for the 100 parameter sets for each validation year. We analysed the validation results of all years combined and for all nine combinations of dry, mean and wet years separately.

Because the focus of this study was on the value of limited inaccurate streamflow observations for model calibration, i.e. the difference in the performance of the models calibrated with the synthetic data series compared to the performance of the models calibrated with hourly FOEN data, all model validation performances are expressed relative to the average $P_{OA}$ of the model calibrated with the hourly FOEN data (our upper benchmark, representing the fully informed case when continuous high quality streamflow data are available). A relative $P_{OA}$ of 1 indicates that the model performance is as good as the performance of the model calibrated with the hourly FOEN data, whereas lower $P_{OA}$ values indicate a poorer performance.

In humid climates, the input data (precipitation and temperature) often dictate that model simulations can not be too far off as long as the water balance is respected (Seibert et al., 2018). To assess the value of limited inaccurate streamflow data for model calibration compared to a situation without any streamflow data, a lower benchmark (Seibert et al., 2018) was used. Here, the lower benchmark was defined as the median performance of the model ran with 1000 random parameters sets. By running the model with 1000 randomly chosen parameter sets, we represent a situation where no streamflow data for calibration are available and the model is driven only by the temperature and precipitation data. We used 1000 different parameter sets to cover most of the model variability due to the different parameter combinations. The Mann–Whitney U test was used to evaluate whether the median $P_{OA}$ for a specific error group and temporal resolution of the data was significantly different from the median $P_{OA}$ for the lower benchmark (i.e. the model runs with random parameters). We furthermore checked for differences in model performance for models calibrated with the same data errors but different temporal resolutions using a Kruskal–Wallis test. By applying a Dunn–Bonferroni post hoc test (Bonferroni, 1936; Dunn, 1959, 1961), we analysed which of the validation results were significantly different from each other.

The random generation of the 100 crowdsourced-like datasets (i.e. the Crowd52 and Crowd12 scenario) for each of the catchments and year characteristics resulted in time series with a different number of high flow estimates. In order to find out whether the inclusion of more high flow values resulted in a better validation performance, we defined the threshold for high flows as the streamflow value that was exceeded 10 % of the time in the hourly FOEN streamflow dataset. The Crowd52 and Crowd12 datasets were then divided into a group that had more than the expected 10 % high flow observations and a group that had fewer high flow observations. To determine if more high flow data improve model performance, the Mann–Whitney U test was used to compare the relative median $P_{OA}$ of the two groups.

## 3 Results

### 3.1 Upper benchmark results

The model was able to reproduce the measured streamflow reasonably well when the complete and unchanged hourly FOEN datasets were used for calibration, although there were also a few exceptions. The average validation $P_{OA}$ was 0.61 (range: 0.19–0.83; Table 4). The validation performance was poorest for the Guerbe (validation $P_{OA} = 0.19$) because several high flow peaks were missed or underestimated by the model for the wet validation year. Similarly, the validation for the Mentue for the dry validation year resulted in a low $P_{OA}$ (0.23) because a very distinct peak at the end of the year was missed and summer low flows were overestimated. The third lowest $P_{OA}$ value was also for the Guerbe (dry validation year) but already had a $P_{OA}$ of 0.35. Six out of the nine lowest $P_{OA}$ values were for dry validation years. Validation for wet years for the models calibrated with data from wet years resulted in the best validation results (i.e. highest $P_{OA}$ values; Table 4).

### 3.2 Effect of errors on the model validation results

Not surprisingly, increasing the errors in the streamflow data used for model calibration led to a decrease in the model performance (Fig. 4). For the small error category, the median validation performance was better than the lower benchmark for all temporal resolutions (Fig. 4 and Table S2). For the medium error category, the median validation performance was also better than the lower benchmark for all scenarios, except for the Crowd12 dataset. For the model calibrated with the dataset with large errors, only the Hourly dataset was significantly better than the lower benchmark (Table 5).

### 3.3 Effect of the data resolution on the model validation results

The Hourly measurement scenario resulted in the best validation performance for each error group, followed by the Weekly data, and then usually the Crowd52 data (Fig. 4). Although the median validation performance of the models calibrated with the Weekly datasets was better than for the Crowd52 dataset for all error cases, the difference was only statistically significant for the no error category (Fig. 5).

The validation performance of the models calibrated with the Weekly and Crowd52 datasets was better than for the scenarios focused on spring and summer observations (WeekendSpring, WeekendSummer and IntenseSummer). The median model performance for the Weekly dataset was significantly better than the datasets focusing on spring and summer for the no, small and medium error groups. The median performance of the Crowd52 dataset was only significantly better than all three measurement scenarios focusing on spring or summer for the small error case (Fig. 5). The model validation performance for the WeekendSummer and IntenseSummer scenarios decreased faster with increasing errors compared to the Weekly, Crowd52 or WeekendSpring datasets (Fig. 4). The median validation $P_{OA}$ for the models calibrated with the WeekendSpring observations was better than for the models calibrated with the WeekendSummer and IntenseSummer datasets but the differences were only significant for the small, medium and large error groups. The differences in the model performance results for the observation strategies that focussed on summer (IntenseSummer and WeekendSummer) were not significant for any of the error groups (Fig. 5).

The median model performance for the regularly spaced Monthly datasets with 12 observations was similar to the median performance for the three datasets focusing on summer with 46–54 measurements (WeekendSpring, WeekendSummer and IntenseSummer), except for the case of large errors for which the monthly dataset performed worse. The irregularly spaced Crowd12 time series resulted in the worst model performance for each error group, but the difference from the performance for the regularly spaced Monthly data was only significant for the dataset with large errors (Fig. 5).

### 3.4 Effect of errors and data resolution on the parameter ranges

For most parameters the spread in the optimised parameter values was smallest for the upper benchmark. The spread in the parameter values increased with increasing errors in the data used for calibration, particularly for MAXBAS (the routing parameter) but also for some other parameters (e.g. TCALT, TT and BETA). However, for some parameters (e.g. CFMAX, FC, and SFCF) the range in the optimised parameter values was mainly affected by the temporal resolution of the data and the number of data points used for calibration. It should be noted though that the changes in the range of model parameters differed significantly for the different catchments and the trends were not very clear.

**Table 4.** Median and the full range of the overall consistency performance $P_{OA}$ scores for the upper benchmark (hourly FOEN data). The $P_{OA}$ values for the dry, average and wet calibration years were used as the upper benchmarks for the evaluation based on the year character (Figs. 6 and S2 in the Supplement); the values in the "overall median" column were used as the benchmarks in the overall median performance evaluation shown in Fig. 4.

| Calibration year | Dry | Average | Wet | Overall median |
|---|---|---|---|---|
| Validation wet year | | | | |
| Upper benchmark | 0.63<br>(0.19–0.79) | 0.65<br>(0.36–0.8) | 0.66<br>(0.45–0.8) | |
| Lower benchmark | | 0.34<br>(−0.02–0.47) | | Upper benchmark |
| Validation average year | | | | 0.61 |
| Upper benchmark | 0.59<br>(0.49–0.64) | 0.61<br>(0.45–0.78) | 0.53<br>(0.36–0.77) | (0.19–0.83) |
| Lower benchmark | | 0.36<br>(0.03–0.59) | | Lower benchmark<br>0.34 |
| Validation dry year | | | | (−0.02–0.59) |
| Upper benchmark | 0.51<br>(0.35–0.71) | 0.59<br>(0.41–0.83) | 0.53<br>(0.23–0.74) | |
| Lower benchmark | | 0.35<br>(0.09–0.52) | | |

### 3.5 Influence of the calibration and validation year and number of high flow data points on the model performance

The influence of the validation year on the model performance was larger than the effect of the calibration year (Figs. 6 and S2). In general model performance was poorest for the dry validation years. The model performances of all datasets with fewer observations or bigger errors than the Hourly datasets without errors were not significantly better than the lower benchmark for the dry validation years, except for Crowd52 in the no error group when calibrated with data from a wet year. However, even for the wet validation years some observation scenarios of the no error and small error group did not lead to significantly better model validation results compared to the median validation performance for the random parameters. Interestingly, the IntenseSummer dataset in the no error group resulted in a very good performance when the model was calibrated for a dry year and also validated in a dry year compared to its performance in the other calibration and validation year combinations. The median model performance was however not significantly better than the lower benchmark due to the low performance for the Guerbe and Allenbach (outliers beyond figure margins in Fig. 6). The validation results for these two catchments were the worst for all the no error–IntenseSummer datasets for all calibration and validation year combinations.

For 13 out of the 18 catchment and year combinations, the Crowd52 datasets with fewer than 10 % high streamflow data points led to a better validation performance than the Crowd52 datasets with more high streamflow data points. For six of them, the difference in model performance was significant. For none of the five cases where more high flow data points led to a better model performance was the difference significant. Also when the results were analysed by year character or catchment, there was no improvement when more high flow values were included in the calibration dataset.

## 4 Discussion

### 4.1 Usefulness of inaccurate streamflow data for hydrological model calibration

In this study, we evaluated the information content of streamflow estimates by citizen scientists for calibration of a bucket-type hydrological model for six Swiss catchments. While the hydroclimatic conditions, the model or the calibration approach might be different in other studies, these results should be applicable for a wide range of cases. However, for physically based spatially distributed models that are usually not calibrated automatically, the use of limited streamflow data would probably benefit from a different calibration approach. Furthermore, our results might not be applicable in arid catchments where rivers become dry for some periods of the year because the linear reservoirs used in the HBV model are not appropriate for such systems.

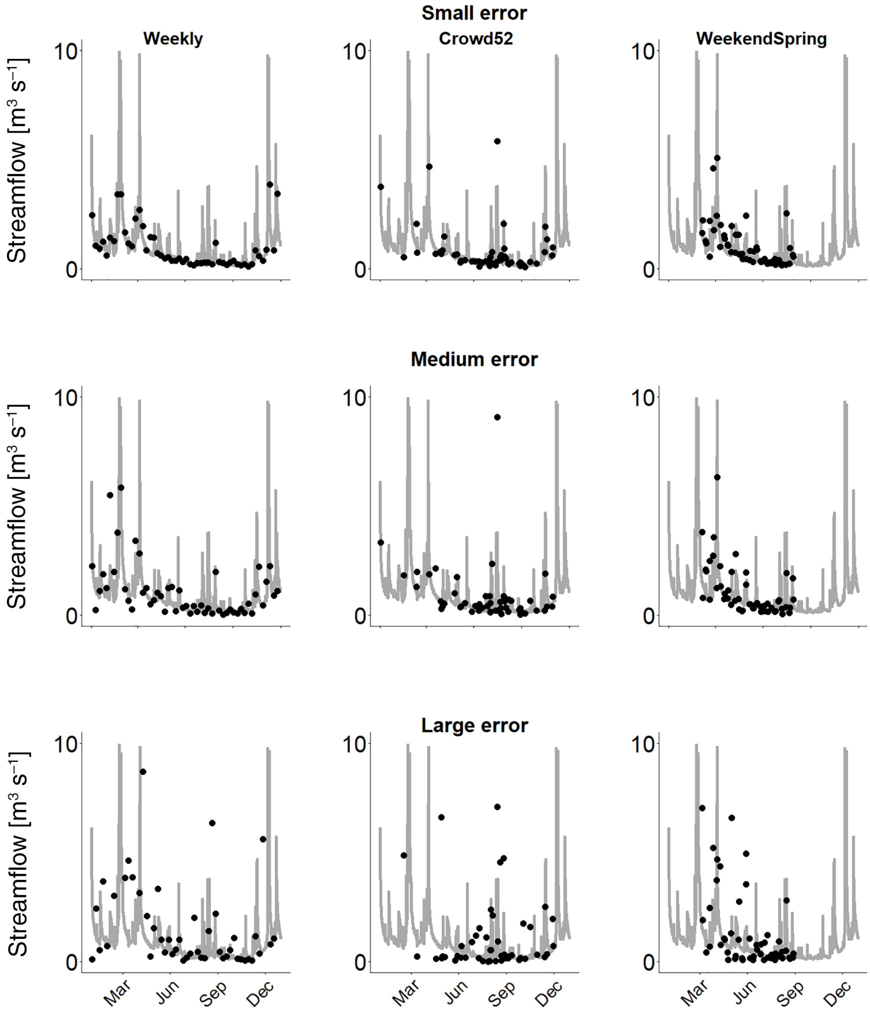



**Figure 3.** Examples of streamflow time series used for calibration with small, medium and large errors and different temporal resolutions (Weekly, Crowd52 and WeekendSpring) for the Mentue in 2010. Large error: adjusted FOEN data with errors resulting from the log-normal distribution fitted to the streamflow estimates from citizen scientists (see Fig. 2). Medium error: same as large error, but the standard deviation of the log-normal distribution was divided by 2. Small error: same as the large error, but the standard deviation of the log-normal distribution was divided by 4. The grey line represents the measured streamflow, and the dots the derived time series of streamflow observations. Note that especially in the large error category some dots lie outside the figure margins.

Streamflow estimates by citizens are sometimes very different from the measured values, and the individual estimates can be disinformative for model calibration (Beven, 2016; Beven and Westerberg, 2011). The results show that if the streamflow estimates by citizen scientists were available at a high temporal resolution (hourly), these data would still be informative for the calibration of a bucket-type hydrological model despite their high uncertainties. However, observations with such a high resolution are very unlikely to be obtained in practice. All scenarios with error distributions that represent the estimates from citizen scientists with fewer observations were no better than the lower benchmark (using random parameters). With medium errors, however, and one data point per week on average or regularly spaced monthly data, the data were informative for model parameterisation. Reducing the standard deviation of the error distribution by a factor of 4 led to a significantly improved model performance compared to the lower benchmark for all the observation scenarios.

A reduction in the errors of the streamflow estimates could be achieved by training of citizen scientists (e.g. videos), improved information about feasible ranges for stream depth, width and velocity, or examples of streamflow values for well-known streams. Filtering of extreme outliers can also reduce the spread of the estimates. This could be done with existing knowledge of feasible streamflow values for a catchment of a given area or the amount of rainfall right before the estimate is made to determine if streamflow is likely to be higher or lower than for the previous estimate. More de-

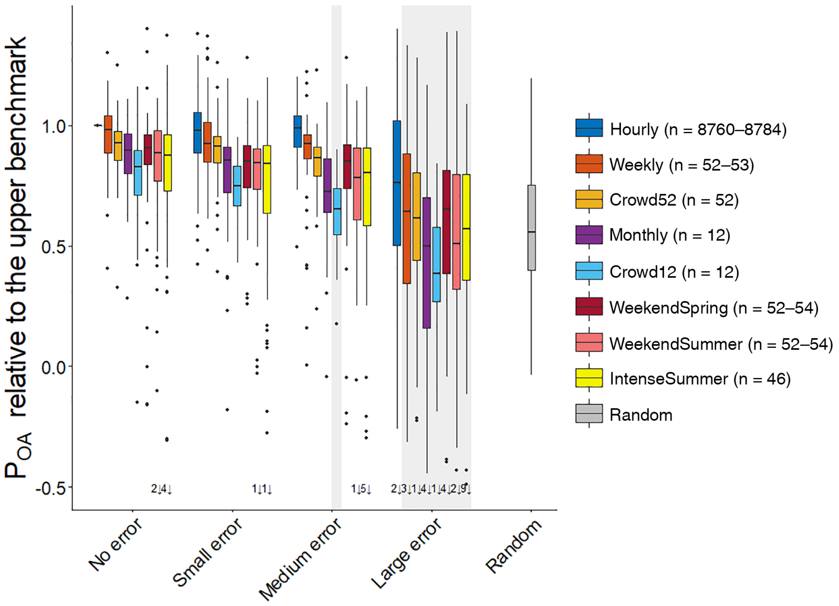



**Figure 4.** Box plots of the median model performance relative to the upper benchmark for all datasets. The grey rectangles around the boxes indicate non-significant differences in median model performance compared to the lower benchmark with random parameter sets. The box represents the 25th and 75th percentile, the thick horizontal line represents the median, the whiskers extend to 1.5 times the interquartile range below the 25th percentile and above the 75th percentile and the dots represent the outliers. The numbers at the bottom indicate the number of outliers beyond the figure margins; *n* is the number of streamflow observations used for model calibration. The result of the hourly benchmark FOEN dataset has some spread because the results of the 100 parameters sets were divided by their median performance. A relative $P_{OA}$ of 1 indicates that the model performance is as good as the performance of the model calibrated with the hourly FOEN data (upper benchmark).

tailed research is necessary to test the effectiveness of such methods.

Le Coz et al. (2014) reported an uncertainty in stage–discharge streamflow measurements of around 5 %–20 %. McMillan et al. (2012) summarised streamflow uncertainties from stage–discharge relationships in a more detailed review and gave a range of ±50 %–100 % for low flows, ±10 %–20 % for medium or high (in-bank) flows and ±40 % for out-of-bank flows. The errors for the most extreme outliers in the citizen estimates are considerably higher, and could differ up to a factor of 10 000 from the measured value in the most extreme but rare cases (Fig. 2). Even with reduced standard deviations of the error distribution by a factor of 2 or 4, the observations in the most extreme cases can still differ by a factor of 100 and 10. The percentage of data points that differed from the measured value by more than 200 % was 33 % for the large error group, 19 % for the medium error group and 4 % for the small error group. Only 3 % of the data points were more than 90 % below the measured value in the large error group and 0 % for both in the medium and small error classes. If such observations are used for model calibration without filtering, they are seen as extreme floods or droughts, even if the actual conditions may be close to average flow. Beven and Westerberg (2011) suggest isolating periods of disinformative data. It is therefore beneficial to identify such extreme outliers, independent of a model, e.g. with knowledge of feasible maximum and minimum streamflow quantities, as used in this study, with the help of the maximum regionalised specific streamflow values for a given catchment area.

### 4.2 Number of streamflow estimates required for model calibration

In general, one would assume that the calibration of a model becomes better when there are more data (Perrin et al., 2007), although others have shown that the increase in model performance plateaus after a certain number of measurements (Juston et al., 2009; Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015). In this study, we limited the length of the calibration period to 1 year because in practice it may be possible to obtain a limited number of measurements during a 1-year period for ungauged catchments before the model results are needed for a certain application, as has been assumed in previous studies (Pool et al., 2017; Seibert and McDonnell, 2015). While a limited number of observations (12) was informative for model calibration when the data uncertainties were limited, the results of this study also suggest that the performance of bucket-type models decreases faster with increasing errors when fewer data points are available (i.e. there was a faster decline in model performance with in-

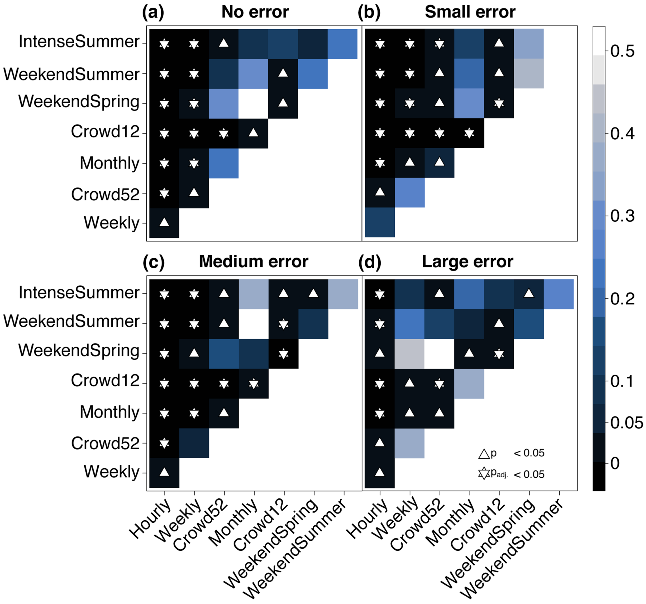



**Figure 5.** Results ($p$ values) of the Kruskal–Wallis with Bonferroni post hoc test to determine the significance of the difference in the median model performance for the data with different temporal resolutions within each data quality group (no error **a**, small error **b**, medium error **c**, and large error **d**). Blue shades represent the p values. White triangles indicate $p$ values $< 0.05$ and white stars indicate p values that, when adjusted for multiple comparisons, are still $< 0.05$.

creasing errors for models calibrated with 12 data points than for the models calibrated with 48–52 data points). This finding was most pronounced when comparing the model performance for the small and medium error groups (Fig. 4). These findings can be explained by the compensating effect of the number of observations and their accuracy because the random errors for the inaccurate data average out when a large number of observations are used, as long as the data do not have a large bias.

### 4.3 Best timing of streamflow estimates for model calibration

The performance of the parameter sets depended on the timing and the error distribution of the data used for model calibration. The model performance was generally better if the observations were more evenly spread throughout the year. For example, for the cases of no and small errors, the performance of the model calibrated with the Monthly dataset with 12 observations was better than for the IntenseSummer and WeekendSummer scenarios with 46–54 observations. Similarly, the less clustered observation scenarios performed better than the more clustered scenarios (i.e. Weekly vs. Crowd52, Monthly vs. Crowd12, Crowd52 vs. IntenseSummer, etc.). This suggests that more regularly distributed data over the year lead to a better model calibration. Juston et al. (2009) compared different subsamples of hydrological data for a 5.6 km$^2$ Swedish catchment and found that including inter-annual variability in the data used for the calibration of the HBV model reduced the model uncertainties. More evenly distributed observations throughout the year might represent more of the within-year streamflow variability and therefore result in improved model performance. This is good news for using citizen science data for model calibration as it suggests that the timing is not as important as the number of observations because it is likely much easier to get observations throughout the year than during specific periods or flow conditions.

When comparing the WeekendSpring, WeekendSummer and IntenseSummer datasets, it seems that it was in most cases more beneficial to include data from spring rather than summer. This tendency was more pronounced with increasing data errors. The reason for this might be that the WeekendSpring scenario includes more snowmelt or rain-on-snow event peaks, in addition to usually higher baseflow, and therefore contains more information on the inter-annual variability in streamflow.

By comparing different variations of 12 data points to calibrate the HBV model, Pool et al. (2017) found that a dataset that contains a combination of different maximum (monthly, yearly etc.) and other flows in model calibration led to the best model performance but also that the differences in performance for the different datasets covering the range of flows were small. In our study we did not specifically focus on the high or low flow data points, and therefore did not have datasets that contained only high flow estimates, which would be very difficult to obtain with citizen science data. However, our findings similarly show that for model calibration for catchments with seasonal variability in streamflow it is beneficial to obtain data for different magnitudes of flow. Furthermore, we found that data points during relatively dry periods are beneficial for validation or prediction in another year and might even be beneficial for years with the same characteristics, as was shown for the improved validation performance of the IntenseSummer dataset compared to the other datasets when data from dry years were used for calibration (Fig. 6).

### 4.4 Effects of different types of years on model calibration and validation

The calibration year, i.e. the year in which the observations were made, was not decisive for the model performance. Therefore, a model calibrated with data from a dry year can still be useful for simulations for an average or wet year. This also means that data in citizen science projects can be collected during any year and that these data are useful for simulating streamflow for most years, except the driest years. However, model performance did vary significantly for the different validation years. The results during dry validation years were almost never significantly better than the

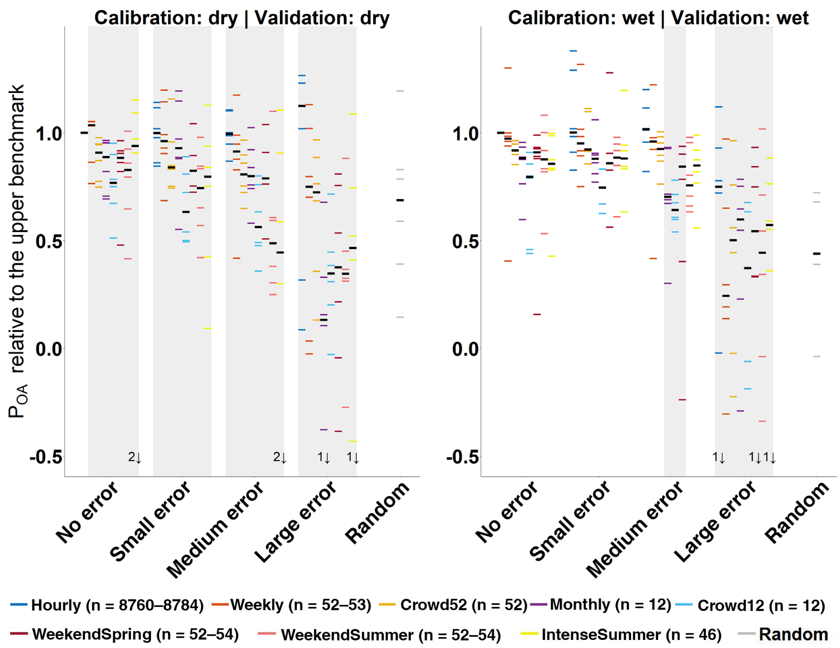



**Figure 6.** Median model validation performance for the datasets calibrated and validated both in a dry year and in a wet year. Each horizontal line represents the median model performance for one catchment. The black bold line represents the median for the six catchments. The grey rectangles around the boxes indicate non-significant differences in median model performance for the six catchments compared to the lower benchmark with random parameters. The numbers at the bottom indicate the number of outliers beyond the figure margins. For the individual $P_{OA}$ values of the upper benchmark (no error–Hourly dataset) in the different calibration and validation years, see Table 4.

lower benchmark (Fig. S2). This might be due to the objective function that was used in this study. Especially the NSE was lower for dry years because the flow variance (i.e. the denominator in the equation) is smaller when there is a larger variation in streamflow. Also, these results are based on six median model performances, and therefore, outliers have a big influence on the significance of results (Fig. S2).

Lidén and Harlin (2000) used the HBV-96 model by Lindström et al. (1997) with changes suggested by Bergström et al. (1997) for four catchments in Europe, Africa and South America. They achieved better model results for wetter catchments and argued that during dry years evapotranspiration plays a bigger role and therefore the model performance is more sensitive to inaccuracies in the simulation of the evapotranspiration processes. The fact that we used a very simple method to calculate the potential evapotranspiration (McGuinness and Bordne, 1972) might also explain why the model performed less well during dry years.

The model parameterisation, obtained from calibration using the IntenseSummer dataset, resulted in a surprisingly good performance for the validation for a more extreme dry year for four out of the six catchments. For the two catchments for which the performance for the IntenseSummer dataset was poor (Guerbe and Allenbach), the weather stations are located outside the catchment boundaries. Especially during dry periods missed streamflow peaks due to misrepresentation of precipitation can affect model performance a lot. The fact that always one of these two catchments had the worst model performance for all the no error–IntenseSummer runs furthermore indicates that the July–September period might not be suitable to represent characteristic runoff events for these catchments. The bad performance for these two catchments for the IntenseSummer–no error run with calibration and validation in the dry year resulted in the insignificant improvement in model performance compared to the lower benchmark. Because the wetness of a year was based on the summer streamflow, these findings suggest that data obtained during times of low flow result in improved validation performance during dry years compared to data collected during other times (Fig. S2). This suggests that if the interest is in understanding the streamflow response during very dry years, it is important to obtain data during the dry period. To test this hypothesis, more detailed analyses are needed.

### 4.5 Recommendations for citizen science projects

Our results show that streamflow estimates from citizens are not informative for hydrological model calibration, unless the errors in the estimates can be reduced through training or advanced filtering of the data to reduce the errors (i.e. to reduce the number of extreme outliers). In order to make streamflow estimates useful, the standard deviation of the error distribution of the estimates needs to be reduced by a factor of 2. Gibson and Bergman (1954) suggest that errors in distance estimates can be reduced from 33 % to 14 % with very little training. These findings are encouraging, although their tests covered distances larger than 365 m (400 yards) and the widths of the medium-sized rivers for which the streamflow was estimated were less than 40 m (Strobl et al., 2018). Options for training might be tutorial videos, as well as lists with values for the width, average depth and flow velocity of well-known streams (Strobl et al., 2018). In order to determine the effect of training on streamflow estimates, further research has to be done because especially the depth estimates were inaccurate (Strobl et al., 2018).

The findings of this study suggest the following recommendations for citizen science projects that want to use streamflow estimates:

- *Collect as many data points as possible*. In this study hourly data always led to the best model performance. It is therefore beneficial to collect as many data points as possible. Because it is unlikely that hourly data are obtained, we suggest to aim for (on average) one observation per week. Provided that the standard deviation of the streamflow estimates can be reduced by a factor of 2, 52 observations (as in the Crowd52 data series) are informative for model calibration. Therefore, it is essential to invest in advertisement of a project and to find suitable locations where many people can potentially contribute, as well as to communicate to the citizen scientists that it is beneficial to submit observations regularly.

- *Encourage observations throughout the year*. To further improve the model performance, or to allow for greater errors, it is beneficial to have observations at all types of flow conditions during the year, rather than during a certain season.

Observations during high streamflow conditions were in most cases not more informative than flows during other times of the year. Efforts to ask citizens to submit observations during specific flow conditions (e.g. by sending reminders to the citizen observers) do not seem to be very effective in light of the above findings. It is rather more beneficial to remind them to submit observations regularly.

Instead of focussing on training to reduce the errors in the streamflow estimates, an alternative approach for citizen science projects is to switch to a parameter that is easier to estimate, such as stream levels (Lowry and Fienen, 2013). Recent studies successfully used daily stream-level data (Seibert and Vis, 2016) and stream-level class data (van Meerveld et al. 2017) to calibrate hydrological models, and other studies have demonstrated the potential value of crowdsourced stream level data for providing information on, e.g. baseflow (Lowry and Fienen, 2013), or for improving flood forecasts (Mazzoleni et al., 2017). However, further research is needed to determine if real crowdsourced stream-level (class) data are informative for the calibration of hydrological models.

## 5 Conclusions

The results of this study extend previous studies on the value of limited hydrological data for hydrological model calibration or the best timing of streamflow measurements for model calibration (Juston et al., 2009; Pool et al., 2017; Seibert and McDonnell, 2015) that did not consider observation errors. This is an important aspect, especially when considering citizen science approaches to obtain streamflow data. Our results show that inaccurate streamflow data can be useful for model calibration, as long as the errors are not too large. When the distribution of errors in the streamflow data represented the distribution of the errors in the streamflow estimates from citizen scientists, this information was not informative for model calibration (i.e. the median performance of the models calibrated with these data was not significantly better than the median performance of the models with random parameter values). However, if the standard deviation of the estimates is reduced by a factor of 2, then the (less) inaccurate data would be informative for model calibration. We furthermore demonstrated that realistic frequencies for citizen science projects (one observation on average per week or month) can be informative for model calibration. The findings of studies such as the one presented here provide important guidance on the design of citizen science projects as well as other observation approaches.

*Data availability.* The data are available from FOEN (streamflow) and MeteoSwiss (precipitation and temperature). The HBV software is available at https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html (Seibert and Vis, 2012) or from jan.seibert@geo.uzh.ch.

*Supplement.* The supplement related to this article is available online at: https://doi.org/10.5194/hess-22-5243-2018-supplement.

*Author contributions.* While JS and IvM had the initial idea, the concrete study design was based on input from all authors. SE and BS conducted the field surveys to determine the typical errors in streamflow estimates. The simulations and analyses were performed by SE. The writing of the manuscript was led by SE; all co-authors contributed to the writing.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* We thank all citizen scientists who participated in the field surveys, as well as the Swiss Federal Office for the Environment for providing the streamflow data, MeteoSwiss for providing the weather data, Maria Staudinger, Jan Schwanbeck and Scherrer AG for the permission to use their datasets and the reviewers for the useful comments. This project was funded by the Swiss National Science Foundation (project CrowdWater).

Edited by: Nadav Peleg
Reviewed by: two anonymous referees

## References

Aschwanden, H. and Weingartner, R.: Die Abflussregimes der Schweiz, Geographisches Institut der Universität Bern, Abteilung Physikalische Geographie, Gewässerkunde, Bern, Switzerland, 1985.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Sveriges Meteorologiska och Hydrologiska Institut (SMHI), Norrköping, Sweden, available at: https://www.researchgate.net/publication/255274162_Development_and_Application_of_a_Conceptual_Runoff_Model_for_Scandinavian_Catchments (last access: 3 October 2018), 1976.

Bergström, S., Carlsson, B., Grahn, G., and Johansson, B.: A More Consistent Approach to Watershed Response in the HBV Model, Vannet i Nord., 4, 1997.

Beven, K.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrol. Sci. J., 61, 1652–1665, https://doi.org/10.1080/02626667.2015.1031761, 2016.

Beven, K. and Westerberg, I.: On red herrings and real herrings: disinformation and information in hydrological inference, Hydrol. Process., 25, 1676–1680, https://doi.org/10.1002/hyp.7963, 2011.

Bonferroni, C. E.: Teoria statistica delle classi e calcolo delle probabilità, st. Super. di Sci. Econom. e Commerciali di Firenze, Istituto superiore di scienze economiche e commerciali, Florence, Italy, 62 pp., 1936.

Brath, A., Montanari, A., and Toth, E.: Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model, J. Hydrol., 291, 232–253, https://doi.org/10.1016/j.jhydrol.2003.12.044, 2004.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., De BiÃv¨re, B., Bhusal, J., Clark, J., Dewulf, A., Foggin, M., Hannah, D. M., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandeya, B., Paudel, D., Sharma, K., Steenhuis, T., Tilahun, S., Van Hecken, G., and Zhumanova, M.: Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development, Front. Earth Sci., 2, 21 pp., https://doi.org/10.3389/feart.2014.00026, 2014.

Davids, J. C., van de Giesen, N., and Rutten, M.: Continuity vs. the Crowd – Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations, Environ. Manage., 60, 12–29, https://doi.org/10.1007/s00267-017-0872-x, 2017.

Davids, J. C., Rutten, M. M., Shah, R. D. T., Shah, D. N., Devkota, N., Izeboud, P., Pandey, A., and van de Giesen, N.: Quantifying the connections – linkages between land-use and water in the Kathmandu Valley, Nepal, Environ. Monit. Assess., 190, 17 pp., https://doi.org/10.1007/s10661-018-6687-2, 2018.

Dickinson, J. L., Zuckerberg, B., and Bonter, D. N.: Citizen Science as an Ecological Research Tool: Challenges and Benefits, Annu. Rev. Ecol. Evol. Syst., 41, 149–172, https://doi.org/10.1146/annurev-ecolsys-102209-144636, 2010.

Dunn, O. J.: Estimation of the Medians for Dependent Variables, Ann. Math. Stat., 30, 192–197, https://doi.org/10.1214/aoms/1177706374, 1959.

Dunn, O. J.: Multiple Comparisons among Means, J. Am. Stat. Assoc., 56, 52–64, https://doi.org/10.1080/01621459.1961.10482090, 1961.

Ewen, T., Brönnimann, S., and Annis, J.: An extended Pacific-North American index from upper-air historical data back to 1922, J. Climate, 21, 1295–1308, https://doi.org/10.1175/2007JCLI1951.1, 2008.

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S., and Burlando, P.: The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model, Water Resour. Res., 47, 14 pp., https://doi.org/10.1029/2010WR009824, 2011.

Finger, D., Vis, M., Huss, M., and Seibert, J.: The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments, Water Resour. Res., 51, 1939–1958, https://doi.org/10.1002/2014WR015712, 2015.

Fitzner, D., Sester, M., Haberlandt, U., and Rabiei, E.: Rainfall Estimation with a Geosensor Network of Cars – Theoretical Considerations and First Results, Photogramm. Fernerkun., 2013, 93–103, https://doi.org/10.1127/1432-8364/2013/0161, 2013.

Gibson, E. J. and Bergman, R.: The effect of training on absolute estimation of distance over the ground, J. Exp. Psychol., 48, 473–482, https://doi.org/10.1037/h0055007, 1954.

Haberlandt, U. and Sester, M.: Areal rainfall estimation using moving cars as rain gauges – a modelling study, Hydrol. Earth Syst. Sci., 14, 1139–1151, https://doi.org/10.5194/hess-14-1139-2010, 2010.

Harrelson, C. C., Rawlins, C. L., and Potyondy, J. P.: Stream channel reference sites: an illustrated guide to field technique, Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station location, Fort Collins, CO, US, 1994.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., and Pierrefeu, G.: Impact of Stage Measurement Errors on Streamflow Uncertainty, Water Resour. Res., 54, 1952–1976, https://doi.org/10.1002/2017WR022039, 2018.

Juston, J., Seibert, J., and Johansson, P.: Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment, Hydrol. Process., 23, 3093–3109, https://doi.org/10.1002/hyp.7421, 2009.

Koch, J. and Stisen, S.: Citizen science: A new perspective to advance spatial pattern evaluation in hydrology, PLoS One, 12, 1–20, https://doi.org/10.1371/journal.pone.0178165, 2017.

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., and Le Boursicaud, R.: Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach, J. Hydrol., 509, 573–587, https://doi.org/10.1016/j.jhydrol.2013.11.016, 2014.

Lidén, R. and Harlin, J.: Analysis of conceptual rainfall–runoff modelling performance in different climates, J. Hydrol., 238, 231–247, https://doi.org/10.1016/S0022-1694(00)00330-9, 2000.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Lowry, C. S. and Fienen, M. N.: CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists, Ground Water, 51, 151–156, https://doi.org/10.1111/j.1745-6584.2012.00956.x, 2013.

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., and Solomatine, D. P.: Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?, Hydrol. Earth Syst. Sci., 21, 839–861, https://doi.org/10.5194/hess-21-839-2017, 2017.

McGuinness, J. and Bordne, E.: A comparison of lysimeter-derived potential evapotranspiration with computed values, Agricultural Research Service – United States Department of Agriculture Location, Washington D.C., 1972.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrol. Process., 24, 1270–1284, https://doi.org/10.1002/hyp.7587, 2010.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrol. Process., 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Michel, C., Perrin, C., and Andreassian, V.: The exponential store: a correct formulation for rainfall – runoff modelling, Hydrol. Sci. J., 48, 109–124, https://doi.org/10.1623/hysj.48.1.109.43484, 2003.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model?, J. Hydrol., 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Perrin, C., Ouding, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, Hydrol. Sci. J., 52, 131–151, https://doi.org/10.1623/hysj.52.1.131, 2007.

Pool, S., Viviroli, D., and Seibert, J.: Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?, J. Hydrol., 554, 613–622, https://doi.org/10.1016/j.jhydrol.2017.09.037, 2017.

Ruhi, A., Messager, M. L., and Olden, J. D.: Tracking the pulse of the Earth's fresh waters, Nat. Sustain., 1, 198–203, https://doi.org/10.1038/s41893-018-0047-7, 2018.

Scherrer AG: Verzeichnis grosser Hochwasserabflüsse in schweizerischen Einzugsgebieten, Auftraggeber: Bundesamt für Umwelt (BAFU), Abteilung Hydrologie, Reinach, 2017.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, https://doi.org/10.5194/hess-4-215-2000, 2000.

Seibert, J. and Beven, K. J.: Gauging the ungauged basin: how many discharge measurements are needed?, Hydrol. Earth Syst. Sci., 13, 883–892, https://doi.org/10.5194/hess-13-883-2009, 2009.

Seibert, J. and McDonnell, J. J.: Gauging the Ungauged Basin?: Relative Value of Soft and Hard Data, J. Hydrol. Eng., 20, A4014004-1–6, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861, 2015.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.

Seibert, J. and Vis, M. J. P.: How informative are stream level observations in different geographic regions?, Hydrol. Process., 30, 2498–2508, https://doi.org/10.1002/hyp.10887, 2016.

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. J.: Upper and lower benchmarks in hydrological modelling, Hydrol. Process., 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018.

Shiklomanov, A. I., Lammers, R. B., and Vörösmarty, C. J.: Widespread decline in hydrological monitoring threatens Pan-Arctic Research, Eos, Trans. Am. Geophys. Union, 83, 13–17, https://doi.org/10.1029/2002EO000007, 2002.

Sideris, I. V., Gabella, M., Erdin, R., and Germann, U.: Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland, Q. J. Roy. Meteor. Soc., 140, 1097–1111, https://doi.org/10.1002/qj.2188, 2014.

Strobl, B., Etter, S., van Meerveld, I., and Seibert, J.: Accuracy of Crowdsourced Streamflow and Stream Level Class Estimates, Hydrol. Sci. J., (special issue on hydrological data: opportunities and barriers), in review, 2018.

van Meerveld, H. J. I., Vis, M. J. P., and Seibert, J.: Information content of stream level class data for hydrological model calibration, Hydrol. Earth Syst. Sci., 21, 4895–4905, https://doi.org/10.5194/hess-21-4895-2017, 2017.

Vrugt, J. A., Gupta, H. V., Dekker, S. C., Sorooshian, S., Wagener, T., and Bouten, W.: Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model, J. Hydrol., 325, 288–307, https://doi.org/10.1016/j.jhydrol.2005.10.041, 2006.

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., Rufino, M. C., and Breuer, L.: Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring, Sci. Total Environ., 631–632, 1590–1599, https://doi.org/10.1016/j.scitotenv.2018.03.130, 2018.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, J. Hydrol., 181, 23–48, https://doi.org/10.1016/0022-1694(95)02918-4, 1996.

Supplement of Hydrol. Earth Syst. Sci., 22, 5243–5257, 2018
https://doi.org/10.5194/hess-22-5243-2018-supplement


Hydrology and Earth System Sciences
Open Access

*Supplement of*

# Value of uncertain streamflow observations for hydrological modelling

**Simon Etter et al.**

*Correspondence to:* Simon Etter (simon.etter@geo.uzh.ch)

## Supplemental Material

## Model parameters

**Table S1 Parameter ranges used for calibration of the HBV-model**

| Parameter | Description[a] | Unit | Min | Max |
|---|---|---|---|---|
| Rescaling Parameters of Input Data | | | | |
| PCALT | change in precipitation with elevation | % $(100m)^{-1}$ | 5 | 15 |
| TCALT | change in temperature with elevation | °C $(10m)^{-1}$ | 0.5 | 1.5 |
| Snow and ice melt parameters | | | | |
| TT | threshold temperature for liquid and solid precipitation | °C | -3 | 1 |
| CFMAX | degree-day factor | $mmd^{-1}°C^{-1}$ | 0.06 | 10 |
| SFCF | snowfall correction factor | - | 0.4 | 1.6 |
| CFR | refreezing coefficient | - | 0.001 | 0.9 |
| CWH | water holding capacity of the snow storage | - | 0.001 | 0.9 |
| Soil Parameters | | | | |
| PERC | maximum percolation from upper to lower groundwater storage | mm $d^{-1}$ | 0 | 3 |
| UZL | threshold parameter | mm | 0 | 100 |
| K0 | storage (or recession) coefficient 0 | $d^{-1}$ | 0.001 | 0.5 |
| K1 | storage (or recession) coefficient 1 | $d^{-1}$ | 0.0001 | 0.2 |
| K2 | storage (or recession) coefficient 2 | $d^{-1}$ | 2E-06 | 0.005 |
| MAXBAS | length of triangular weighting function | H | 1 | 7 |
| FC | maximum soil moisture storage | Mm | 50 | 550 |
| LP | soil moisture value above which actual evapotranspiration reaches potential evapotranspiration | - | 0.3 | 1 |
| Beta | shape factor for the function used to calculate the distribution of rain and snow melt going to runoff and soil box, respectively | - | 1 | 5 |

[a]a detailed description of the model parameters is given in (Seibert and Vis, 2012).

## Significance of median model performance compared to the lower benchmark

**Table S2 Significance of the differences in median model performance for each temporal resolution and an error group compared to the lower benchmark (Mann-Whitney U-test). The p-values of the Kruskal-Wallis test for the within group variability in the lowermost row shows that the median model performance of the different error groups was significantly different.**

| | No Error | Small Error | Medium Error | Large Error |
|---|---|---|---|---|
| Hourly | <0.01 | <0.01 | <0.01 | <0.01 |
| Weekly | <0.01 | <0.01 | <0.01 | 0.75 |
| Crowd52 | <0.01 | <0.01 | <0.01 | 0.40 |
| Monthly | <0.01 | <0.01 | <0.01 | 0.03* |
| Crowd12 | <0.01 | <0.01 | 0.11 | <0.01* |
| WeekendSpring | <0.01 | <0.01 | <0.01 | 0.40 |
| WeekendSummer | <0.01 | <0.01 | <0.01 | 0.46 |
| IntenseSummer | <0.01 | 0.01 | 0.04 | 0.21 |
| Within error group | <0.01 | <0.01 | <0.01 | <0.01 |

* These datasets result in significantly worse results than random parameters.

12 **Extreme outlier removal for the northern and southern side of the Alps**

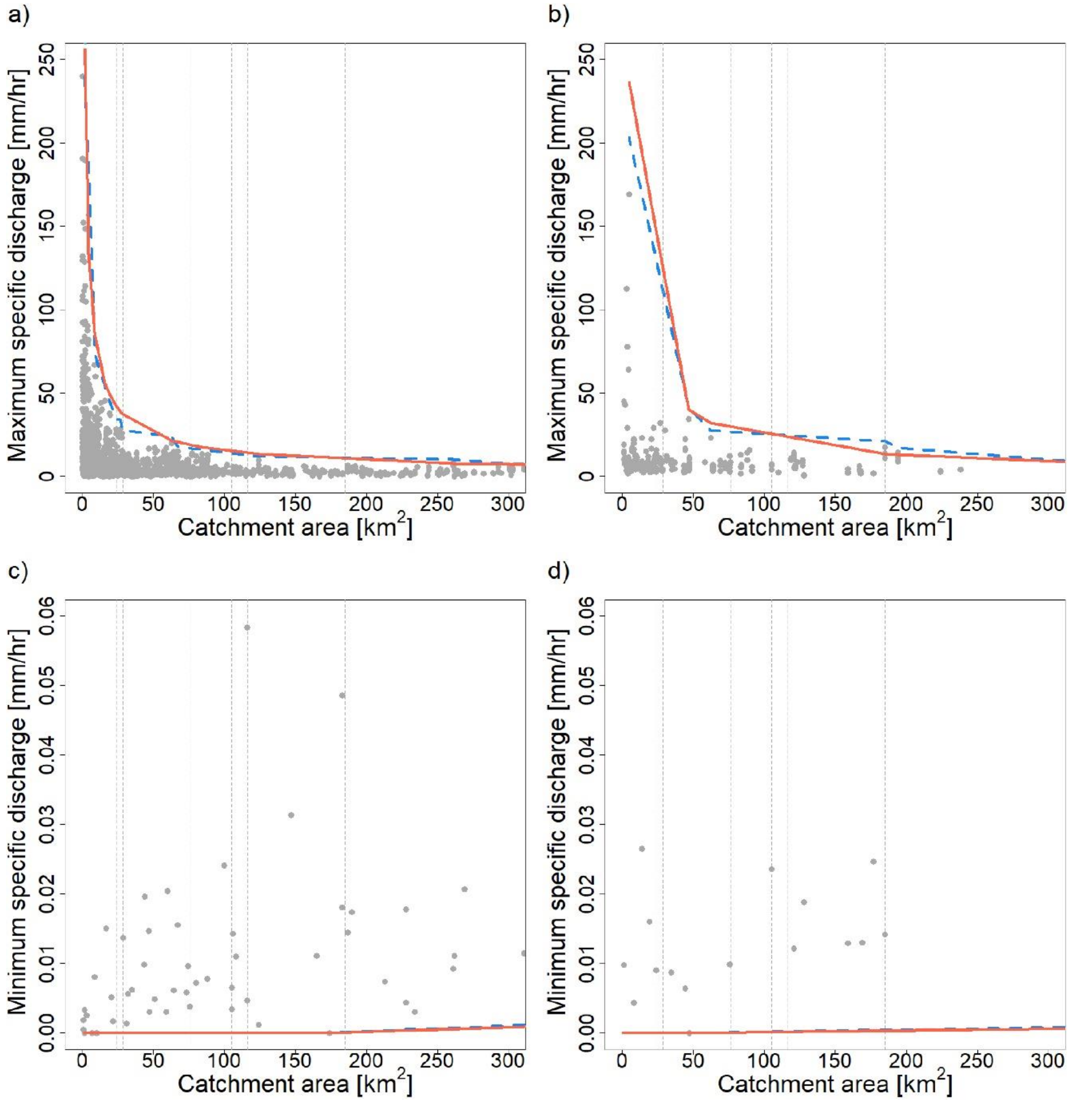



**Figure S1 Relation between catchment area and maximum (a, b) and minimum (c, d) specific streamflow for catchments on the north (a, c) and south (b, d) of the Alps. The dashed light blue line is the Pareto front including the 20 % buffer. The red lines are the fitted logarithmic models used to find the maximum and minimum possible flow for each catchment.**

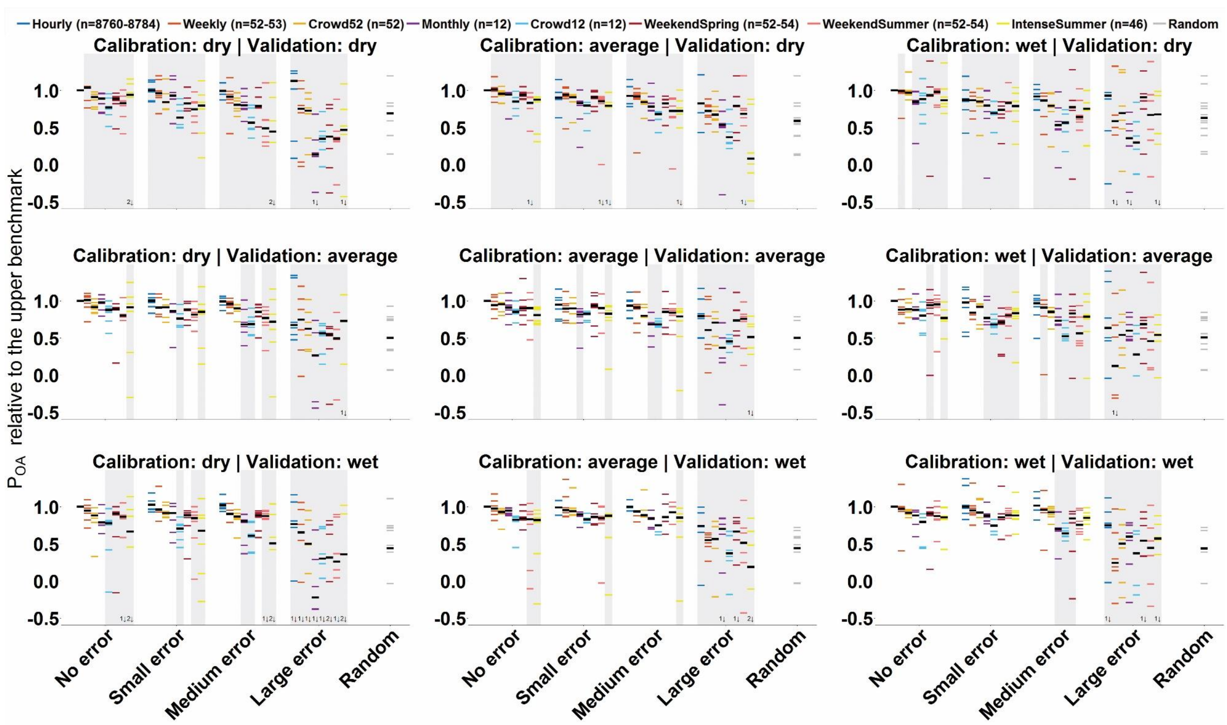



**Figure S2 Median model validation performance for all datasets used for calibration during the different validation periods. Each horizontal line represents the median model performance for one catchment. The black bold line represents the median for the six catchments. The grey rectangles around the boxes indicate non-significant differences in median model performance for the six catchments compared to the lower benchmark with random parameters. The numbers at the bottom indicate the number of outliers beyond the figure margins. For the individual $P_{OA}$ values of the upper benchmark (no error – *Hourly* dataset) in the different calibration and validation years see Table 4.**

# Paper VI

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026108

# Value of Crowd-Based Water Level Class Observations for Hydrological Model Calibration

S. Etter[1], B. Strobl[1], J. Seibert[1,2], and H. J. Ilja van Meerveld[1]

[1]Department of Geography, University of Zurich, Zurich, Switzerland, [2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

**Key Points:**

- Water level class observations can be informative for hydrological model calibration
- Model parameters calibrated with water level class data performed similarly well as those calibrated with precise water level measurements
- Errors in water level class data observations had a minimal effect on the streamflow simulations

**Correspondence to:**
S. Etter,
simon.etter@geo.uzh.ch

**Citation:**
Etter, S., Strobl, B., Seibert, J., & van Meerveld, H. J. I. (2020). Value of crowd-based water level class observations for hydrological model calibration. *Water Resources Research*, *56*, e2019WR026108. https://doi.org/10.1029/2019WR026108

Received 19 AUG 2019
Accepted 25 DEC 2019
Accepted article online 3 JAN 2020



**Abstract** While hydrological models generally rely on continuous streamflow data for calibration, previous studies have shown that a few measurements can be sufficient to constrain model parameters. Other studies have shown that continuous water level or water level class (WL-class) data can be informative for model calibration. In this study, we combined these approaches and explored the potential value of a limited number of WL-class observations for calibration of a bucket-type runoff model (HBV) for four catchments in Switzerland. We generated synthetic data to represent citizen science data and examined the effects of the temporal resolution of the observations, the numbers of WL-classes, and the magnitude of the errors in the WL-class observations on the model validation performance. Our results indicate that on average one observation per week for a 1-year period can significantly improve model performance compared to the situation without any streamflow data. Furthermore, the validation performance for model parameters calibrated with WL-class observations was similar to the performance of the calibration with precise water level measurements. The number of WL-classes did not influence the validation performance noticeably when at least four WL-classes were used. The impact of typical errors for citizen science-based estimates of WL-classes on the model performance was small. These results are encouraging for citizen science projects where citizens observe water levels for otherwise ungauged streams using virtual or physical staff gauges.

**Plain Language Summary** Normally, multiple years of streamflow measurements are used to calibrate a hydrological model for a specific catchment so that it can be used to, for instance, predict floods or droughts. Taking these measurements is expensive and requires a lot of effort. Therefore, such data are often missing, especially in remote areas and developing countries. We investigated the potential value of water level class (WL-class) data for model calibration. WL-classes can be observed by citizens with the help of a virtual ruler with different classes that is pasted onto a picture of a stream bank as a sticker (see Figure 2). We show that one WL-class observation per week for 1 year improves model calibration compared to situations without streamflow data. The model results for the WL-class observations were as good as precise water level observations that require a physical staff gauge or continuous water level data measurements that can be obtained from a water level sensor that is installed in the stream. However, the results were not as good as when streamflow data were used for model calibration, but these are more expensive to collect. Errors in the WL-class observations did in most cases not affect the model performance noticeably.

## 1. Introduction

Hydrological models are usually calibrated with continuous streamflow data acquired at gauging stations. Such data sets are scarce, especially for remote regions and developing countries, even though people in these areas are often affected by various kinds of water issues (Mulligan, 2013). Globally, hydrological observation networks are on the decline, mainly due to reduced financial resources (Kundzewicz, 1997). Furthermore, access to available data is often restricted (Fekete et al., 2012). To collect data in ungauged basins, citizen science approaches that use modern communication technology (i.e., smartphones) can be helpful. Citizen science approaches can also incorporate local knowledge, for instance, for hazard assessment (Sy et al., 2018) and help to raise public awareness of environmental issues (Lanfranchi et al., 2014). However, the usefulness of citizen science data is often questioned due to the perceived lack of experience of the volunteers (Cohn, 2008) and potential biases, such as location bias related to the population density or temporal bias related to the timing of the observations (Kosmala et al., 2016). It is important to standardize measurement protocols (Dickinson et al., 2012), e.g., by using smartphone applications, to evaluate the accuracy and value of the collected data, and to improve the measurement protocols iteratively when needed. It is also useful to thoroughly examine the potential use of citizen science data before starting a new project.

Publications that include citizen science projects focusing on water quantity in streams are still rather scarce; most publications on water related citizen science projects have focused on water quality (Buytaert et al., 2014; Njue et al., 2019). Some recent examples of water quantity-focused projects are the EU-funded citizen observatories that aim to complement data collection by authorities, such as WeSenseIt (www.wesenseit.com; Lanfranchi et al., 2014), GroundTruth2.0 (https://gt20.eu), and SCENT (https://scent-project.eu). Projects that specifically focus on streamflow or water levels are CrowdHydrology in the United States (Lowry et al., 2019; Lowry & Fienen, 2013), Smartphones4Water in Nepal (www.smartphones4water.org; Davids et al., 2017), a project in Kenya (www.uni-giessen.de/hydro/hydrocrowd_kenya; Weeser et al., 2018), Cithyd in Italy (www.cithyd.com; Balbo & Galimberti, 2016), and CrowdWater (www.crowdwater.ch; Seibert, Strobl, et al., 2019). The CrowdWater project aims to explore the value of citizen science data and to collect water level class (WL-class) data (Seibert, Strobl, et al., 2019), as well as qualitative data on soil moisture and the state of temporary streams (Kampf et al., 2018; Seibert, van Meerveld, et al., 2019), and riverine export of macro plastic. For observations of WL-classes, virtual staff gauges with class markings are inserted onto a photograph of the streambank, bridge pillar, or other features in the stream. These features and the virtual staff gauge then serve as a reference to which later observations of the water level are compared. Repeated observations result in time series of WL-classes. However, these series are irregular in time and potentially contain observation errors (Strobl et al., 2019a).

Several studies have examined the value of discontinuous streamflow data for the calibration of hydrological models. For example, Pool et al. (2019) investigated the value of a limited number of streamflow measurements for calibration of the HBV model (Bergström, 1976; Lindström et al., 1997) and found that 12 measurements taken during a 1-year period can lead to satisfying model simulations. Seibert and McDonnell (2015) showed for the Maimai catchment in New Zealand that streamflow measurements throughout an event or 10 observations during high flow periods provide as much information for model calibration as 3 months of continuous measurements. These model studies assumed error-free streamflow measurements. All measurements are affected by errors, and these can be considerable for streamflow measurements (particularly during high flows or low flows; McMillan et al., 2018), but for citizen science data, errors might be particularly large (Aceves-Bueno et al., 2017). This can significantly limit the value of the data. Therefore, we previously investigated the value of streamflow data that included errors that are typical for citizen-based estimates of streamflow (Etter et al., 2018). We found that streamflow estimates from citizens, who did not receive any form of training, did not improve model performance compared to a model with random parameter sets. We concluded that either the errors in the streamflow estimates have to be reduced by some form of training or that a quantity, that is easier to estimate, such as water levels or WL-classes, should be used (Strobl et al., 2019a). Water level measurements require the installation of a staff gauge. Citizens then can read the water level from the staff gauge and report them via text messages or a smartphone application. Previous studies have shown that this method works well and can provide useful and accurate data (Lowry et al., 2019; Weeser et al., 2018). However, the installation of a staff gauge can be complicated in practice. Beyond issues such as how to securely fix the gauge, permissions by local authorities might be required. Obtaining permits can require time and effort and cause additional costs. WL-class estimates, as used within the CrowdWater project, do not require a physical staff gauge and are, thus, more scalable. However, the data have a lower precision (and likely also lower accuracy) than readings from a staff gauge.

Continuous (e.g., daily) water level or WL-class data can be informative for hydrological model calibration. Seibert and Vis (2016) concluded that the use of daily water level data for model calibration results in a surprisingly good model performance, especially for humid catchments. For arid regions additional information was necessary to achieve a good simulation. In another study, van Meerveld et al. (2017) showed that daily WL-class data are informative for hydrological model calibration as well, and that the performance of the model calibrated with WL-class data with at least five equally frequent classes was not much worse than a model calibrated with water level data.

We aim to develop a methodology that is quick and easy to use for citizen scientists, while at the same time being robust and informative for the calibration of hydrological models and to thereby extend the knowledge on the potential of crowdsourced data with different qualities as proposed in Weeser et al. (2019). We therefore investigated the potential value of discontinuous WL-class data as these can be obtained by citizens using synthetic data, which is a similar approach as in Etter et al. (2018). Our objectives were to (i) assess the potential value of a few WL-class observations at intervals that are realistic for citizen science projects, for model calibration; (ii) assess the potential effect of likely errors in WL-class observations on model performance; and (iii) investigate the influence of the number of WL-classes in combination with different observation scenarios on model performance.

## 2. Methods

At the time of writing this paper, an insufficient number of repeated observations had been collected with the CrowdWater App to determine the value of WL-class data for model calibration. We, therefore, used synthetic data (cf. Etter et al., 2018; Seibert & Vis, 2016; van Meerveld et al., 2017), which is an efficient approach to assess data requirements before making considerable efforts to collect the data (Christophersen et al., 1993; Pool et al., 2019). First, we converted the water level time series for four Swiss catchments into WL-class time series. From these continuous data sets, we created time series with fewer data points representing different observation scenarios and introduced errors that are typical for citizen estimates of WL-classes (Strobl et al., 2019a). We then used these synthetic data sets to calibrate a simple bucket-type model, the HBV model (Bergström, 1976; Lindström et al., 1997; Seibert & Vis, 2012). Finally, we used the calibrated parameter sets to evaluate the model performance for the validation period by comparing it to the observed streamflow. We compared the validation performance to the validation performance of the model calibrated with the original (continuous, and assumed to be error free) streamflow data (upper benchmark), and the validation performance of the noninformed case, where the model is run with random parameter sets (lower benchmark).

### 2.1. Catchments

For this study, we selected four gauged catchments in Switzerland with different flow regimes (Aschwanden & Weingartner, 1985). Streamflow measurements at the outlet of these catchments have good quality for both high and low flow conditions and are unaffected by backwater issues. Furthermore, the catchments are relatively little affected by anthropogenic influences and have no glaciers. The catchment areas range from 79 to 186 km$^2$ and the mean elevations range from 652 to 1,651 m a.s.l. (Table 1 and Figure 1).

### 2.2. HBV Model

We used the bucket-type hydrological model HBV (Lindström et al., 1997), which was originally developed at the Swedish Meteorological and Hydrological Institute (SMHI) by Bergström (1976). The HBV model consists of routines for snow storage, soil water, and groundwater. In this study, we used the model implementation HBV-light (Seibert & Vis, 2012). The catchments were divided into elevation zones, each covering a band of 100 m, for which the snow, soil, and groundwater routines were computed individually.

### 2.3. Measured Data

Water level and streamflow time series were obtained from the Swiss Federal Office for the Environment (FOEN). The 10-min measurements were averaged to obtain hourly water level and streamflow time series. Hourly areal precipitation sums were obtained from the CombiPrecip data set of MeteoSwiss (Sideris et al., 2014). The data for the years 2011 and 2013 suggest an unrealistic high runoff-rainfall ratio (>0.9) for the Verzasca catchment and were, thus, excluded from all simulations. A possible reason is that the weather stations are located outside the catchment and that precipitation is highly variable in this alpine terrain. Furthermore, the station data used in the CombiPrecip data set are not corrected for wind undercatch, which can lead to errors of up to 40% in winter for windy locations in Switzerland (Sevruk, 1985).

The hourly temperature at the mean elevation of the catchment was calculated from data from nearby weather stations (see Table 1 and Figure 1) using Thiessen polygons and a lapse rate of −6°C per 1,000 m. Data gaps existed only in the hourly temperature datasets. The most extended gaps covered 5 days and were filled with interpolated data. The potential evapotranspiration was calculated using the day of the year, the latitude, and the temperature following the approach of McGuinness and Bordne (1972). We chose this simple model because more physically based potential evapotranspiration models would require more input data, which are not available with a satisfying spatial resolution in alpine terrain.

**Table 1**
*Catchment Characteristics for the Four Swiss Catchments Used in This Study*

| Catchment | | Murg | Guerbe | Mentue | Verzasca |
|---|---|---|---|---|---|
| Gauging station (FOEN station number) | | Waengi (2126) | Belp, Mülimatt (2159) | Yvonand, La Mauguettaz (2369) | Lavertezzo, Campiòi (2605) |
| Weather stations | | Aadorf-Taenikon, Hörnli | Plaffeien, Bern-Zollikofen | Mathod, Pully | Acquarossa, Cimetta, Magadino, Piotta |
| Area [$km^2$] | | 79 | 117 | 105 | 186 |
| Elevation [m a.s.l.] | Min | 465 | 522 | 445 | 490 |
| | Max | 1,035 | 2,176 | 927 | 2,864 |
| Regime Type[a] | | Pluvial-inférieur | Pluvial-supérieur | Pluvial-jurassien | Nivo-pluvial-méridional |
| Min/Max Pardé coefficients | | 0.68/1.34 | 0.77/1.39 | 0.46/1.57 | 0.23/2.22 |
| Mean annual streamflow Q [mm/y] | | 756 | 746 | 491 | 1,764 |
| Mean annual precipitation P [mm/y] | | 1,343 | 1,319 | 1,287 | 2,014 |
| Mean runoff ratio (Q/P) | | 0.56 | 0.57 | 0.38 | 0.88 |
| July–September streamflow [mm] (calibration\|validation) | | | | | |
| Dry | | 90\| 86 | 106\| 94 | 26\| 24 | 324\| 307 |
| Average | | 125\| 149 | 202\| 195 | 54\| 62 | 417\| 439 |
| Wet | | 220\| 228 | 308\| 451 | 93\| 187 | 670\| 810 |
| Annual runoff ratio (calibration\| validation) | | | | | |
| Dry | | 0.72\| 0.54 | 0.37\| 0.82 | 0.41\| 0.41 | 0.98[b]\| 0.71 |
| Average | | 0.55\| 0.43 | 0.48\| 0.60 | 0.52\| 0.65 | 0.66\| 0.63 |
| Wet | | 0.56\| 0.54 | 0.54\| 0.81 | 0.50\| 0.52 | 1.32[b]\| 0.73 |

*Note*. Long-term annual averages were computed for the period 1974–2014, except for Verzasca for which the 1990–2014 period was used.
[a]Regime types according to Aschwanden and Weingartner (1985). [b]For Verzasca the calibration years 2011 and 2013 have an unrealistic runoff-rainfall ratio (>0.9) and were therefore excluded from all simulations (see text).

### 2.4. Selection of Years for Model Calibration and Validation

To obtain information on the influence of wetness conditions on the value of citizen science-derived WL-class data for model calibration, we selected for each catchment an average, a dry and a wet year for model calibration and validation. For the average year, we selected the two years within the 2006–2014 period (the period with available hourly precipitation data at the time of the study) for which the total summer streamflow (July–September) was closest to the average summer streamflow for the 1974–2014 period. For the wet and the dry year, we selected the two years with the highest and lowest streamflow sum during the summer, respectively. For the calibration, we used the years that were second closest to the average, highest, or lowest value; for the validation, we used the year that were closest to the average and the years with the highest and lowest total streamflow during the summer (Table 1). Even though citizen science projects can obtain long-term data (e.g., the Audubon Christmas Bird Count has collected data for more than 100 years; Meehan et al., 2019), we wanted to test the value of 1 year of citizen science-derived WL-class data for hydrological modeling because in reality most studies do not have time to obtain more extended time series.

### 2.5. Synthetic Data

#### 2.5.1. WL-Class Time Series

We assume that the WL-class observations are made at the catchment outlet. In order to determine the effect of the number of classes, we split the water level records from the FOEN into 2 to 10, 15, and 20 classes, resulting in 11 different WL-class time series per catchment. The WL-classes could, for instance, be obtained from a photograph of the stream with a sticker of a staff gauge added to it. The case with 10 classes corresponds to the "virtual staff gauge" approach used in the CrowdWater app (Seibert, Strobl, et al., 2019; see example in Figure 2). The class borders were set at equal water level intervals between the fifth and 95th percentile of the water level record for the period for which the rating curve did not change and included the calibration years (Table 2). The cumulative frequency distribution of the water levels was approximately linear between the fifth and 95th percentile for all four catchments. Water levels below the fifth and above the 95th percentile would likely be below or above the virtual staff gauges set by the citizen scientists and were assigned to the lowest and highest WL-classes, respectively (Figure 4).

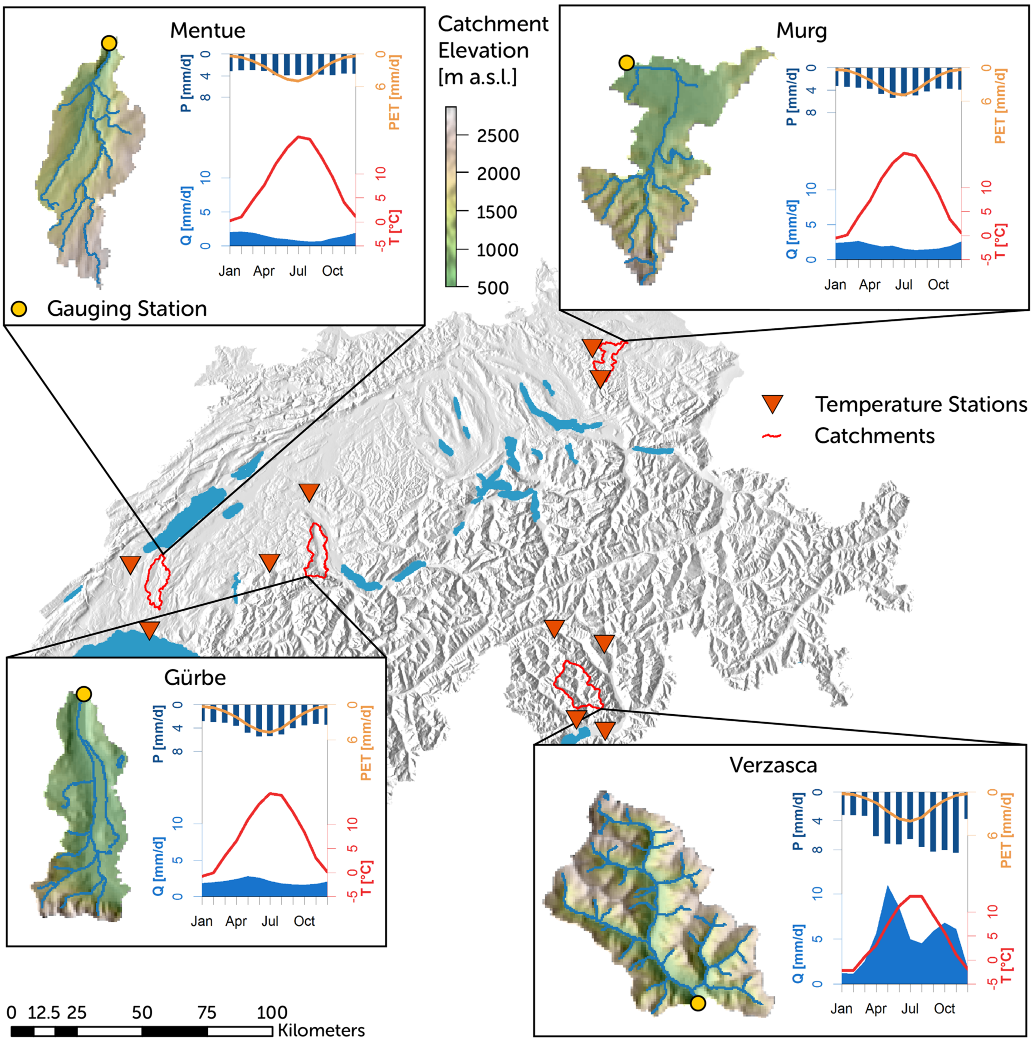



**Figure 1.** Map of Switzerland showing the location of the four catchments and the weather stations used to derive the temperature data. For each catchment, monthly average precipitation (P), streamflow (Q), temperature (T), and potential evapotranspiration (PET) are shown for the period 1974–2014, except for Verzasca for which the period 1990–2014 was used.

### 2.5.2. Observation Scenarios

We created water level and WL-class time series for observation scenarios that differed in the number of observations and the clustering of the observations throughout the year (Table 3). We used the same observation scenarios as Etter et al. (2018) for comparability. For the *Crowd52* and *Crowd12* scenarios, we assigned higher probabilities to periods when people are more likely to be outdoors (i.e., a higher probability for summer than winter, a higher probability for weekends than weekdays, and a higher probability outside office hours; see Table 3 in Etter et al., 2018). This led to a larger number of observations during the summer for the *Crowd52* scenario than the *Weekly* scenario (median of 33 observations between May and September for *Crowd52* vs. 22 for *Weekly*) and for *Crowd12* vs. the *Monthly* data (median of 8 for *Crowd12* vs. 5 for *Monthly*). In citizen science projects, the number of contributions will vary but based on our experience in the CrowdWater project, we assume that these scenarios cover a wide range of plausible cases.

In addition to the scenarios of Etter et al. (2018), we added the daily resolution for comparability with the results of van Meerveld et al. (2017). Daily data are not likely for citizen science projects but near-daily data are possible: In CrowdHydrology 347 observations per year were made in the location with most contributions (Lowry et al., 2019). The location with most contributions in CrowdWater receives on average one

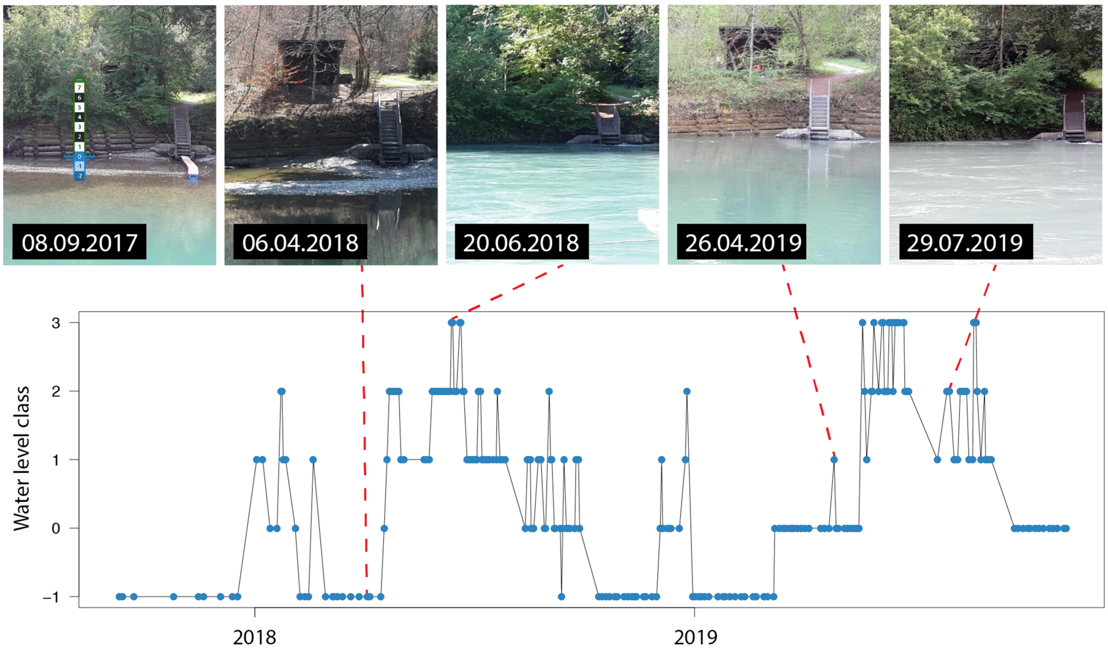



**Figure 2.** Time series of WL-class observations at the Aare river in Zollikofen, Switzerland, based on the virtual staff gauge inserted on the reference picture (left picture in the upper row of the figure), which can then be used to estimate the water level class at the later dates (other pictures in the upper row). The entire time series of observations for this location can be found online (https://www.spotteron.com/crowdwater/spots/141766). Note that this time series illustrates the water level class data that can be observed by citizen scientists; we did not use this time series in the modeling described in this study. All photos were taken by Auria Buchs.

observation every 1.2 days and for the location shown in Figure 2 there was on average one observation every 3.2 days. The hourly water level data represent data from a water level logger, while hourly WL-class data could potentially be obtained from webcam images.

#### 2.5.3. Adding Errors to the WL-Class Time Series With 10 Classes

Citizen science-derived data likely contain errors. We assessed the typical errors in WL-class observations in a series of field surveys (Strobl et al., 2019a). We analyzed 440 estimates of WL-classes from citizens who compared the water level in the stream that they were looking at to a photo of the same stream taken at an earlier time with a sticker of a staff gauge with 10 classes added to it (the first photo in Figure 2 shows

**Table 2**
*The Time Periods of the Water Level Records That Were Used to Determine the WL-Class Boundaries and the Dry, Average, and Wet Years Chosen for Model Calibration and Validation*

| | Murg | Guerbe | Mentue | Verzasca |
|---|---|---|---|---|
| Period used for class definition | 1974–2014 | 1996–2009[a] | 1974–2014 | 1990–2013 |
| Calibration years | | | | |
| Dry | 2013 | 2011a | 2010 | 2013 |
| Average | 2008 | 2008 | 2006 | 2007 |
| Wet | 2007 | 2007 | 2014 | 2011 |
| Validation years | | | | |
| Dry | 2009 | 2013 | 2009 | 2010 |
| Average | 2011 | 2006 | 2013 | 2006 |
| Wet | 2014 | 2014 | 2007 | 2008 |

*Note*. The rating curves did not change considerably during the selected time period to determine the WL-class boundaries.
[a]For the Guerbe catchment, the dry calibration (2011) year occurred in a period after the rating curve changed so that there was a systematic shift in the water level data. Therefore, the class borders were determined for this period separately. For the validation period, we used streamflow data that were calculated with an adapted rating curve and therefore did not include this shift.

**Table 3**
*The Different Scenarios for the Temporal Resolution of the Observations Used in This Study, With the Number of Data Points in 1 Year of Data (n)*

| | |
|---|---|
| *Hourly* | One data point per hour (8,760 ≤ n ≤ 8,784, depending on the year) |
| *Daily* | One data point every day (365 ≤ *n* ≤ 366), randomly between 6 am and 8 pm |
| *Weekly* | One data point per week, every Saturday, randomly between 6 am and 8 pm (52 ≤ *n* ≤ 53) |
| *Monthly* | One data point per month on the 15th of the month, randomly between 6 am and 8 pm (*n* = 12) |
| *IntenseSummer* | One data point every other day between July and September, randomly between 6 am and 8 pm (~15 observations per month, *n* = 46) |
| *WeekendSummer* | One data point each Saturday and each Sunday between May and October, randomly between 6 am and 8 pm (52 ≤ *n* ≤ 54) |
| *WeekendSpring* | One data point on each Saturday and each Sunday between March and August, randomly between 6 am and 8 pm (52 ≤ *n* ≤ 54) |
| *Crowd52* | 52 data points (in order to be comparable to the *Weekly*, *IntenseSummer*, and *WeekendSpring* time series), between 6 am and 8 pm |
| *Crowd12* | 12 data points (comparable to the *Monthly* data), between 6 am and 8 pm |

an example). Nearly half (48%) of the participants chose the right class (as determined by experts) and 40% were off by only one class (Strobl et al., 2019a). The errors (i.e., the difference between the reported WL-class and the actual WL-class as determined by experts) were approximately normally distributed (Figure 3). We used these discrete class error probabilities to add random errors to each WL-class data point for the scenarios with 10 WL-classes (Figure 4). The same probability of errors was used for all four watersheds and years. In addition to this error, hereafter referred to as large error, we also created two time series with reduced errors to consider possible benefits of training or error-filtering (e.g., via reassessment of the WL-class data by multiple volunteers based on a comparison of images; Strobl et al., 2019b):

- Large error: Typical errors of citizen scientists, i.e., random errors according to the normal distribution of errors from the survey of Strobl et al. (2019), as shown in Figure 3.
- Medium error: Random errors according to the normal distribution with the standard deviation divided by two.
- Small error: Random errors according to the normal distribution with the standard deviation divided by four.
- No error: The 10 classes based on water level measurements by the FOEN, which are considered to be error-free and the benchmark in terms of quality for WL-class data.

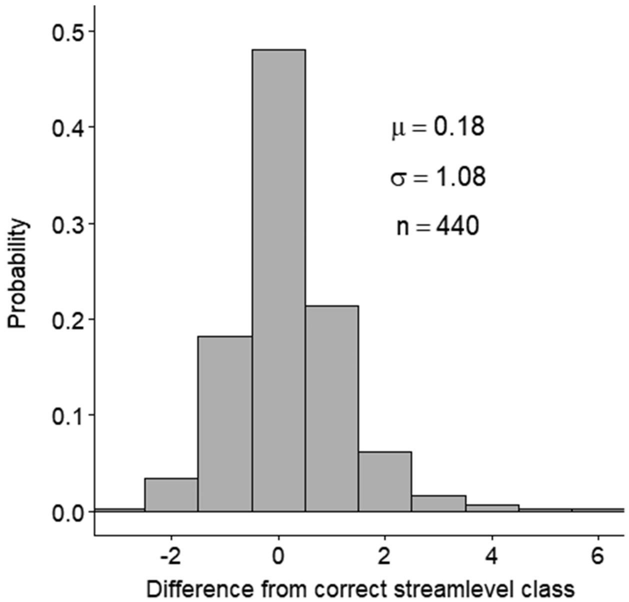



**Figure 3.** Distribution of the errors in the WL-class estimates (i.e., the difference between the reported WL-class and the actual WL-class, as determined by experts) from field surveys for nine different locations. The data were obtained from Strobl et al. (2019a). This distribution was used to create WL-class time series with large errors.

### 2.6. Model Calibration

We calibrated the hydrological model for each of the synthetic data series (nine different temporal resolutions, three error magnitudes with 10 classes, and 11 class sizes without errors) for each of the three calibration years for each of the four catchments. We also calibrated the model for the nine different temporal resolutions of the water level data and the hourly streamflow data for each year and catchment. For the calibration with measured streamflow, we used the overall performance index ($P_{OA}$; Finger et al., 2011). The $P_{OA}$ is the mean of the Nash-Sutcliffe efficiency for the streamflow (Nash & Sutcliffe, 1970), the Nash-Sutcliffe efficiency for the log-transformed streamflow, the mean absolute relative error, and the volume error. For each calibration with water level or WL-class data, we optimized the Spearman rank correlation coefficient (Spearman, 1904) for the relation between the synthetic WL-class data and the simulated streamflow using a genetic optimization algorithm (Seibert, 2000). The calibration ranges for the 16 parameters were based on their typical range and are the same as in Etter et al. (2018). For each calibration, we used the preceding year as the warm-up period and calibrated the model 100 times to account for parameter uncertainty. Each model calibration consisted of 3,500 model runs and 1,000 runs for local optimization. This resulted in 100 parameter sets for each of the three hourly streamflow calibrations (dry, average, and wet year, respectively), each of the 27 water level simulations (3 years and nine temporal resolutions), and each of the 378 WL-class simulations

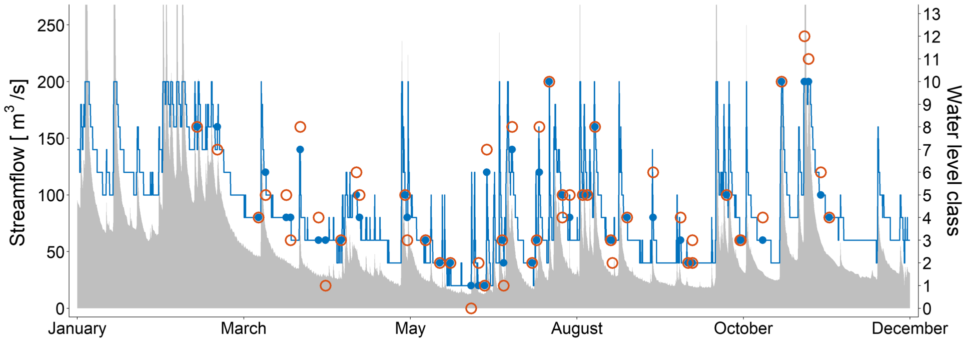



**Figure 4.** Observed streamflow at Mentue in 2014 (gray area), the hourly WL-class time series with 10 classes (blue line) derived from continuous water level data, and the synthetic data series for the *Crowd52* scenario without any errors (blue dots) and large errors (orange circles) that were used for model calibration. The error distribution and formula used to add errors to the WL-classes derived from the water level data are given in Figure 3.

(3 years, nine temporal scenarios, and three error magnitudes plus 11 different class sizes) per catchment, except for Verzasca for which only the average year was used for calibration (Table 1). For the *Crowd52* and *Crowd12* data sets different realizations of the observation times are possible and we, thus, randomly selected different observation times for each of the 100 calibration trials. For these cases, the spread of the results is, thus, a combination of parameter uncertainty and observation timing.

The Spearman rank coefficient cannot be computed if the WL-class data set contained data for only one class (i.e., due to a lack of variation in the water level data). This occurred for less than 1% of all the scenarios studied here. For computation of the Spearman rank coefficient for these scenarios, the WL-class for the observation at the time of the highest streamflow was manually changed to the next (higher) class.

### 2.7. Model Validation

For each scenario, we used the 100 calibrated parameter sets to simulate the streamflow for the validation years. The validation performance was assessed using the overall performance index $P_{OA}$, as was done for the assessment of the value of uncertain streamflow data by Etter et al. (2018). We determined the median of the 100 $P_{OA}$ values for each scenario and compared it to the median $P_{OA}$ of the validation for the model calibrated with the observed streamflow data, which was considered the best possible model performance and thus the upper benchmark.

We similarly compared the median model validation performance for the different WL-class scenarios to the median validation performance of the model calibrated with the hourly water level time series. For each WL-class scenario, we also compared the validation performance to the validation performance of the model calibrated with water level data with the same temporal resolution in order to compare the value of citizen science-based WL-class data and citizen science-based water level data for model calibration. We used the

**Table 4**
*Overview of the Different Model Validation Comparisons Used to Evaluate the Value of Crowdsourced WL-Class Data*

| Validation performance for calibration using WL-class data vs. | Statistically significant difference in median $P_{OA}$ value indicates: |
|---|---|
| Hourly streamflow data (upper benchmark) | A gauging station is more useful for model calibration than citizen science-derived WL-class data using a virtual staff gauge |
| Hourly water level data | Installation of a water level recorder is more useful for model calibration than a virtual staff gauge that citizen scientists can use to determine the WL-class |
| Water level scenarios | Installation of a staff gauge from which citizens can read water levels is more useful for model calibration than a virtual staff gauge to determine the WL-class |
| Random parameter sets (lower benchmark) | Citizen science-derived WL-class data have added value for model calibration |

*Note*. For each comparison the median validation performances were compared using the one-sided paired Wilcoxon test. Significant differences are indicated by filled squares in Figures 5 and 6.

**Table 5**
*Median Validation Performance (i.e., Median $P_{OA}$ Values for the 100 Parameters) for the Different Calibration and Validation Years When the Model was Calibrated With Hourly Streamflow Data (Upper Benchmark)*

| Validation | Dry | | | Average | | | Wet | | | Median |
|---|---|---|---|---|---|---|---|---|---|---|
| Calibration | Dry | Average | Wet | Dry | Average | Wet | Dry | Average | Wet | (of all year combinations) |
| Murg | 0.71 | 0.76 | 0.74 | 0.58 | 0.59 | 0.56 | 0.79 | 0.78 | 0.80 | 0.74 |
| Guerbe | 0.35 | 0.51 | 0.57 | 0.63 | 0.75 | 0.77 | 0.19 | 0.36 | 0.55 | 0.55 |
| Mentue | 0.40 | 0.41 | 0.23 | 0.64 | 0.64 | 0.75 | 0.66 | 0.65 | 0.73 | 0.64 |
| Verzasca | 0.63[a] | 0.83 | 0.48[a] | 0.52[a] | 0.78 | 0.47[a] | 0.65[a] | 0.80 | 0.68[a] | 0.80 |

[a]These years had a runoff rainfall ratio >0.9 (see Table 1) and were not included in any of the other results.

one-sided paired Wilcoxon test to determine if the median model validation performance for the calibration with WL-class data was significantly worse than the validation performance for the model calibrate with the measured water level data. If there is no significant difference, then more easily scalable methods that do not require the installation of sensors, such as virtual staff gauges, are equally useful for model calibration as physical staff gauges. If the performance is significantly worse, it might be useful to invest in the installation of an actual staff gauge and have citizens report the water level from this staff gauge (Table 4).

The lower benchmark was defined as a situation where no streamflow, water level, or WL-class data are available for model calibration. In wet environments, random parameters can result in surprisingly good model performance as long as the model reproduces the water balance. Therefore, the lower benchmark serves as the minimum model performance that can be expected based on the water balance alone (Seibert et al., 2018). Thus, for the lower benchmark, we used the median performance of 1,000 streamflow time series generated from the precipitation and temperature data in the validation period based on 1,000 parameter sets that were selected randomly from the parameter ranges. We then compared the median validation performance of the models calibrated with streamflow, water level, or WL-class data to the median model validation performance for the 1,000 random parameter sets. We tested whether the median model validation performance of the WL-class scenarios (for all nine calibration and validation year combinations for all four catchments) was significantly better than the median validation performance for the random parameters using the one-sided paired Wilcoxon test. We considered the data set useful for calibration when the median validation performance was significantly better than for the random parameters (Table 4).

To determine the significance of differences in the median validation performance for the different observation scenarios (i.e., different temporal resolutions) but the same number of WL-classes and error category, we used a Kruskal-Wallis test with the Dunn Bonferroni post hoc test with adjusted $p$ values for multiple comparisons (Bonferroni, 1936; Dunn, 1959).

## 3. Results

### 3.1. Model Performance for Calibration Based on Hourly Data

In general, the HBV model was able to reproduce the observed streamflow reasonably well when it was calibrated using the hourly streamflow data (upper benchmark). The median $P_{OA}$ value for these calibrations was 0.82 (range: 0.66–0.88, with the lowest value for the calibration of the Guerbe for a dry year). The simulations for the validation period were not as good with a median $P_{OA}$ of 0.64 (range: 0.19–0.83). The lowest validation $P_{OA}$ value was for the Guerbe catchment when it was calibrated for the dry year and validated for the wet year (Table 5). These years had very different runoff-ratios (0.37 for the dry calibration year and 0.81 for the wet validation year; Table 1). The median validation performance (for all combinations of calibration and validation years) was also worst for the Guerbe catchment ($P_{OA} = 0.55$, range for the other catchments 0.64 to 0.80; Table 5).

The median validation result of all model simulations based on model calibration using hourly water level data (median: 0.52; range: −0.39 to 0.78) was significantly worse than for the calibration with the hourly streamflow data ($p < 0.001$; Figure 5). The use of hourly water level data for model calibration caused the most noticeable decline in the median model validation performance for the Guerbe ($P_{OA}$ relative to the upper benchmark: 0.45, range for the other catchments 0.75–0.92).

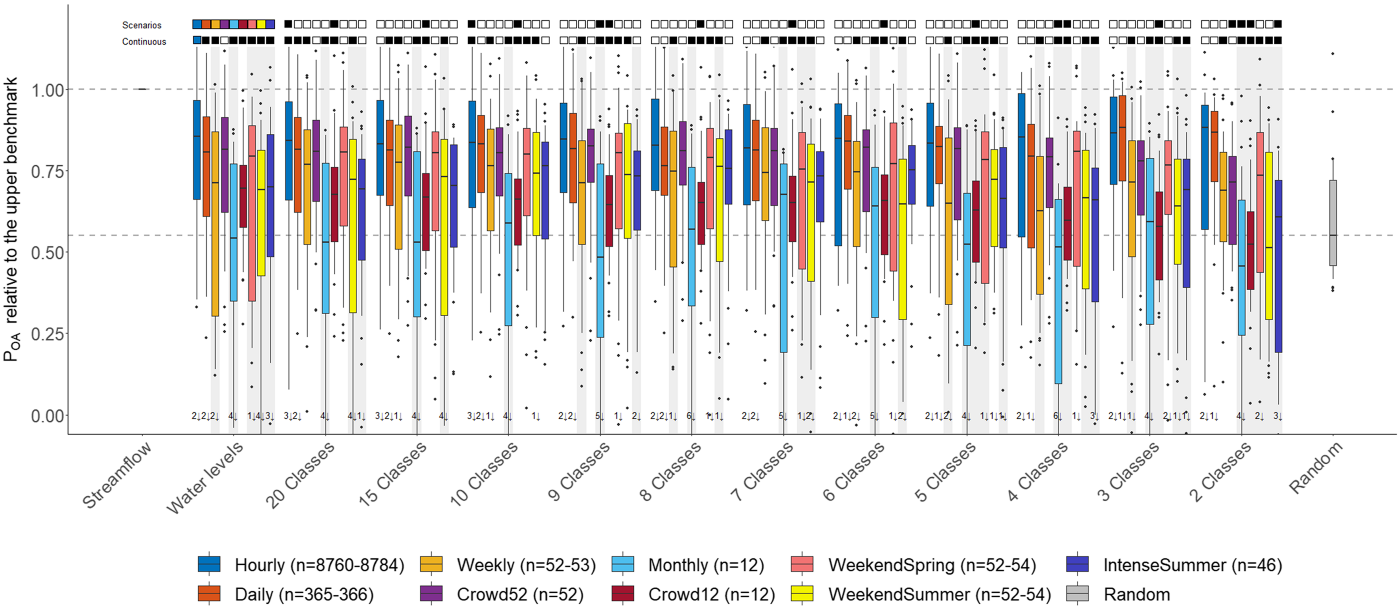



**Figure 5.** Box plots of the validation performance of the HBV-model calibrated with synthetic WL-class data (different temporal resolutions and different numbers of WL-classes) relative to the performance of the model calibrated with hourly streamflow data. The lower benchmark (in gray) represents the median performance of the model run with 1,000 randomly selected parameter sets. The gray background shading highlights the scenarios for which the median model performance was not significantly better than for the lower benchmark. The filled squares at the top of the graph indicate cases where the median validation performance for the model calibrated with WL-class data was significantly worse compared to the calibration with water level data with the same temporal resolution (top row) and compared to the calibration with continuous (hourly) water level data (second row); empty squares indicate no statistically significant difference based on the one-sided paired Wilcoxon test. All scenarios led to a significantly worse model validation performance than calibration with continuous streamflow data. The WL-classes were equally sized and assumed to be error free. The box extends from the 25th to 75th percentile and the whiskers extend to the tenth and ninetieth percentile. The black line inside the box represents the median. Numbers at the bottom indicate outliers with a relative $P_{oa} < 0.00$.

Calibration based on the hourly WL-class data led to a significantly worse median validation performance than calibration using streamflow data, regardless of the number of WL-classes (2–20 classes, all $p < 0.001$). However, for the case without errors, the performance of the model calibrated with hourly WL-class data was not significantly worse than when the hourly water level data (i.e., hourly) were used for calibration, except for the case with 10 classes due to outliers (see white squares in the second row on the top of Figure 5).

### 3.2. Effect of the Number of Observations and WL-Classes (No-Error Case)

In general, the model validation performance was poorer when the model was calibrated with fewer water level or WL-class observations. Overall, the data set with 52 crowdsourcing-like observations (*Crowd52*) led to the best model validation performance of all data sets with on average one observation per week. The scenario with two observations each weekend between March and August (*WeekendSpring*) and the scenario with regularly spaced weekly observations (*Weekly*) led to the next best model performance. Although the median validation performance for the models calibrated with the *WeekendSpring* data was always higher than for the model calibrated with observations each weekend from May to October (*WeekendSummer*), or every other day from July to September (*IntenseSummer*) (Figure 5), this difference was not statistically significant according to the Dunn-Bonferroni test (adjusted $p$ values were all >1.0).

As one would expect, the model validation performance decreased slightly when the number of WL-classes decreased, but the effect depended on the temporal resolution of the data used for calibration (Figure 5). For two WL-classes, only the scenarios *Hourly*, *Daily*, and *Crowd52* led to similar model validation performances as the continuous water level data. For all other scenarios, performances were significantly worse ($p \leq 0.03$).

When daily WL-class data were used, the model validation performance was only for the cases with 15 and 20 classes significantly worse than the performance of the model calibrated with continuous water level data ($p$ values = 0.03 and 0.02, Figure 5). This was largely due to two outliers in both cases in the Guerbe catchment with $P_{OA}$-values between −0.18 and −0.40 or scores relative to the $P_{OA}$ of the upper benchmark between −0.5

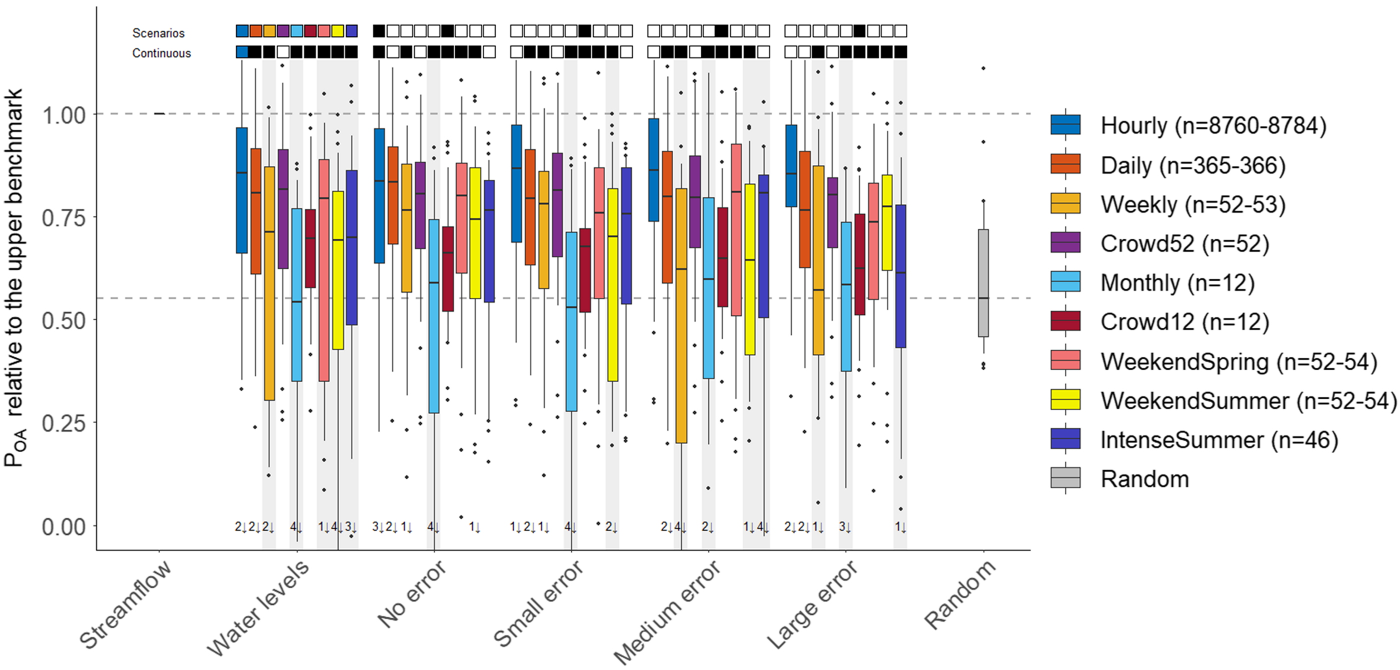



**Figure 6.** Box plots of the validation performance of the HBV-model calibrated with water level data with different temporal resolutions and the synthetic WL-class data (10 equal sized classes) with different temporal resolutions and different errors, relative to the validation performance of the model calibrated with hourly streamflow data (upper benchmark). The lower benchmark shown (in gray) is the median validation performance of the model run with 1,000 random parameters. The gray shading indicates a median model performance that is not significantly better than the lower benchmark ($p > 0.05$). The filled black squares at the top of the graph indicate cases where the median validation performance for the calibration with WL-class data is significantly worse than the calibration with water level data with the same temporal resolution (top row) or compared to continuous water level data (second row); empty squares indicate no statistically significant difference based on the one-sided paired Wilcoxon test.

and −1.9. The validation performance of the model calibrated with the temporally discontinuous *Crowd52* water level or WL-class data sets was never significantly worse than the validation performance of the model calibrated with the continuous water level data. The validation performance of the scenario focused on summer (*IntenseSummer*) was not significantly worse than the validation performance of continuous water level data if five or more WL-classes were used. The median model validation performance for the scenario with two observations each weekend between March and August (*WeekendSpring*) with 3, 4, 6, 15, and 20 WL-classes was not significantly different ($p$ values: 0.05–0.08) to the performance of the model calibrated with the hourly water level data either. This was also the case for the observations every other day between July and September (*IntenseSummer*) with at least five WL-classes ($p$ values: 0.06–0.27). For all the other scenarios, the model validation performance was significantly worse than for the calibration with continuous water level data (see black squares in the second row above the main plot in Figure 5).

Calibration with discontinuous WL-class data led in only very few cases to a significantly poorer model performance than calibration with temporally discontinuous but precisely measured water level data: the *Crowd12* scenario regardless of the number of WL-classes; the *Monthly* scenario with 2, 4, and 9 classes; and the *Crowd52*, *WeekendSpring*, *WeekendSummer*, and *IntenseSummer* scenario with two classes (see black squares in first row above Figure 5).

The validation performance for the model calibrated with the *Hourly*, *Daily*, and *Crowd52* data sets was always better than the lower benchmark. However, monthly WL-class data (*Monthly*) never improved the validation performance compared to the lower benchmark (Figure 5). For five or fewer classes, there were more scenarios for which the model did not perform significantly better than the lower benchmark, e.g., the *Weekly*, *Crowd12*, *WeekendSpring*, *WeekendSummer*, and *IntenseSummer* scenarios. However, the $p$ values were close to 0.05 and therefore the significance test results differed for the different number of WL-classes. The model performance for the *IntenseSummer* and *WeekendSpring* scenarios did not systematically improve with an increasing number of classes, hence model performance for these scenarios was best when eight or nine WL-classes were used.

### 3.3. Effect of Errors in WL-Class Estimates With 10 Classes

Including errors in the WL-class data resulted in only a minor decrease in the overall model validation performance. This effect was particularly small compared to the effect of the temporal resolution of the data used for model calibration (Figure 6). For all *Hourly*, *Daily*, *Crowd52*, *Crowd12*, and the *WeekendSpring* cases with 10 WL-classes, the model validation performance was better than the lower benchmark, even when large errors were included in the calibration data (Figure 6). The effect of errors on model validation performance was most substantial for calibration with the *Weekly* and *IntenseSummer* data sets for which the scenarios with medium and large errors were not significantly better than the lower benchmark. The addition of medium or large errors also caused the validation performance for the model calibrated with the *IntenseSummer* data to become significantly worse than the model calibrated with continuous hourly water level data (Figure 6). The performance of the model calibrated with the *Daily* data became only significantly worse than the model calibrated with continuous water level data when small or medium errors were included. The model validation performance for calibration with *Hourly* and *Crowd52* WL-class data was not significantly worse than the validation performance for calibration with continuous water levels, even with large errors (Figure 6).

The median validation performance of the model calibrated with the discontinuous WL-class data remained similar to the performance of the model calibrated with discontinuous water level data, except for *Crowd12* and *Hourly* WL-classes (again due to the large outliers in the Guerbe catchment) for which the calibration with WL-class data with errors led to a significantly worse validation performance than calibration with discontinuous water level data.

### 3.4. Effects of Variability in WL-Class Data on Model Performance

For the *Crowd52* scenarios, there were 100 realizations for every catchment and year. This allowed us to explore the effect of the distribution of the WL-class observations on model performance. For the wet years with more streamflow in summer, there was a more balanced distribution of data points across the classes than for the dry years. For the wet years, 14% of all data points were in the lowest class, 19% in the second class, and 22% in the third class when 10 WL-classes were used. The corresponding numbers were 20%, 24%, and 17% for the average year and 45%, 16%, and 14% for the dry years. For the *Crowd52* scenario with 10 classes and no errors, the median validation performance for parameter sets obtained by calibration with data from the wet year (median $P_{OA} = 0.54$) and the average year (median $P_{OA} = 0.56$) were significantly better than for the calibration with data from the dry years (median $P_{OA} = 0.44$, both $p \leq 0.001$).

We also compared the model validation performance of *Crowd52* scenarios for WL-class data based on 10 classes and without errors with a different number of observations in classes 1 and 2 (i.e., observations during baseflow conditions). If more than half of the observations were in classes 1 or 2, model validation performance was significantly worse than if there were relatively fewer observations for classes 1 and 2 (and thus more observations for classes 3–10). This indicates that the model performance increases when there are more observations for the higher WL-classes. However, for the *Crowd52* scenario, there was no correlation between the variance in WL-classes used for calibration and model validation performance for the resulting 100 calibrated parameter sets (adjusted coefficients of determination were all $\leq$0.01). This is likely due to the large variability in the individual parameter sets and their effect on model performance because for the *Crowd52* scenario only one parameter set was obtained for each observation scenario.

## 4. Discussion

With this study, we extended our understanding of the value of uncertain data for hydrological model calibration. The usefulness of WL-class data for model calibration was shown earlier for continuous WL-class data for a large number of catchments in the United States by van Meerveld et al. (2017). Here we show that even discontinuous WL-class data are useful for model calibration. We used the HBV model for the analyses but argue that the findings would be similar for other bucket-type hydrological models. For physically based spatially distributed models that are used without calibration, WL-class data might still be useful for model evaluation. The results are likely different for arid regions, where rivers only flow at certain times of the year, as Seibert and Vis (2016) showed that model parameterizations based on calibration against water level data were less suitable to simulate streamflow for arid regions than for humid regions.

### 4.1. Value of WL-Class Data for Hydrological Model Calibration

Usually hydrological models are calibrated using streamflow data derived from water level measurements and a rating curve. In practice, this is the most expensive method to obtain stream-related data but it also leads to the best validation results (which is why we consider this the upper benchmark). Continuous water level measurements are easier to obtain; water level loggers have become cheaper and can now easily store a year of data. However, the installation of water level loggers still requires some investment and maintenance, particularly in steep mountainous terrain where the stream channel may change frequently due to scour and deposition. The different temporal observation scenarios with precise water level data represent the case when a physical staff gauge is placed in a stream and passers-by read the level and transmit their observation, as it is done in the CrowdHydrology project (Lowry & Fienen, 2013), Cithyd (www.cithyd.com; Balbo & Galimberti, 2016), and a project in Kenya (Weeser et al., 2018).

The median validation performance for the model calibrated using WL-class data was worse than for the model that was calibrated using streamflow data but as good as using water level data with the same temporal resolution. Even for the realistic citizen science scenario *Crowd52*, the validation performance was not significantly worse than when hourly water level data that are recorded with a water level logger are used for calibration. These results suggest that while traditional streamflow measurements are most informative for hydrological model calibration and continuous hourly water level data certainly have their value, observations of WL-classes, e.g., based on virtual staff gauges (Seibert, Strobl, et al., 2019), are also valuable for model calibration and can lead to reasonable streamflow simulations when streamflow data are not available. Model calibration with WL-class data can be used to transform the measured WL-classes into streamflow time series and, thus, be used to derive useful information, such as hydrologic signatures (e.g., runoff-rainfall ratios). The use of regionalized parameter values (Andréassian et al., 2014) would be an alternative approach to derive streamflow estimates for ungauged basins. This approach, however, is also subject to uncertainties as the transfer functions will only be approximations (Hundecha et al., 2002). This was, therefore, not part of this study but it raises the interesting question whether a few WL-class observations can improve regionalized parameter sets for areas where there are no other streamflow data.

### 4.2. Effects of Timing of the WL-Class Observations and Errors on Model Performance

The number of observations and the timing of the observations in the year had a larger influence on model performance than errors in the WL-class observations. Errors generally had the largest effect on model performance when few observations were available for calibration, as was the case for the *Monthly* and the *Crowd12* scenarios (Figure 6). Compared to the effect of errors in streamflow estimates on model validation performance (Etter et al., 2018), the effect of errors in the WL-classes was minor. This can be explained by the fact that there are no extreme outliers in the WL-class data and that the errors in the WL-class estimates are smaller than those for streamflow estimates (Strobl et al., 2019a). Even for the large error case, 48% of the observations were in the correct class and 88% of the observations were within ±1 class of the correct class (Strobl et al., 2019; Figure 3).

Although there was a general trend of increasing model performance with an increasing number of observations, the timing of the observations within the year also had a substantial effect on model performance. The validation performance for the model calibrated with *Crowd52* data (i.e., with more observations in summer) was comparable to the performance of the model calibrated with hourly water level data, regardless of the number of classes. On the other hand, the validation performance of the model calibrated with *Weekly* data was significantly worse than the performance of the model calibrated with hourly water level data, even when using 20 WL-classes. This is contrary to the results for uncertain streamflow observations of Etter et al. (2018), where *Weekly* data resulted in a better model validation performance than *Crowd52* data. For WL-class estimates, it is probably beneficial to obtain observations that cover a larger variation in streamflow magnitudes than for streamflow directly because it takes a relatively large change in the actual water level (and thus also streamflow) to change one WL-class. Large variations in streamflow occur more often during wet years with higher flows, leading to the significantly better validation performance for the wet or average years compared to the dry years for the *Crowd52* data set. A denser sampling strategy during summer is also more likely to catch more of the variation in streamflow, leading to the better model validation performance for the model calibrated with *Crowd52* data compared to *Weekly* data and for the *IntenseSummer* data (with observations every other day during June, July, and August) compared to *WeekendSummer* data. This also explains why the *IntenseSummer* scenario led to a similar performance as *WeekendSpring*, even though that scenario covers more streamflow variation during spring. The median model validation performance for the calibration with the *WeekendSpring* data was higher than for the *WeekendSummer* data and comparable (i.e., not significantly worse) to the validation performance of the model calibrated with the hourly water level data, while calibration with the *WeekendSummer* data led to a significantly worse model performance than when hourly water level data were used for calibration.

### 4.3. Influence of the Number of WL-Classes on Model Performance

The staff gauges in the survey of Strobl et al. (2019) had 10 classes and, thus, the errors used in this study are typical for staff gauges with 10 classes. However, in practice fewer classes will be used for many locations as the virtual staff gauges that are inserted in the pictures are often too large, so that it is unlikely that the water level will reach the highest classes (Seibert, Strobl, et al., 2019). Our results indicate, however, that even when the water level fluctuates in only two or three classes, such data can be informative for model calibration if there is on average at least one observation per week. The influence of errors on such observations might, however, be larger than when all 10 classes are used (although the chances for large observation errors are likely smaller for very large virtual staff gauges).

The benefit of using more than four to five WL-classes (depending on the scenario) for model calibration was negligible. This is roughly in line with the findings of van Meerveld et al. (2017), who showed for continuous WL-class data for about 600 catchments in the United States that there was hardly any improvement in model performance if more than five WL-classes were used. However, the observation scenario affects this result, e.g., for the *Weekly* scenario the results tended to be more stable when 10 classes or more were used (Figure 5). However, the results of the scenarios with observations during summer (*WeekendSummer* and *IntenseSummer*) suggest that in terms of model performance it is not necessarily helpful to have more WL-classes. Especially in summer, when extended periods of low flows can be expected, eight to 10 classes might provide the model enough degrees of freedom to perform well in a validation year that is different from the calibration year, whereas more WL-classes can lead to overfitting of the model to the calibration period. During periods of low flow, the water level will vary across more classes when more classes are used (and individual WL-classes are thus smaller), which might lead to overfitting of the model for that particular year and result in calibrated parameter sets that do not perform well during other years.

### 4.4. Implications for Citizen Science Projects

For citizen science projects, where the data quality often is an important issue (Show, 2015), clear and straightforward procedures help to ensure good data quality (Cohn, 2008). Based on the results of this study, a simple approach using a virtual staff gauge with 10 classes (as implemented in the CrowdWater app; Seibert, Strobl, et al., 2019) can provide data that are useful for model calibration. The WL-class estimates seem to be superior for citizen science projects than streamflow estimates as indicated by the significantly better model performance of the *Crowd52* and *WeekendSpring* data sets compared to the calibration using random parameters, even when the errors in the observations were large. This was not the case for streamflow estimates, for which large errors hampered the usefulness of the data for model calibration (Etter et al., 2018).

The lack of an increase in model performance for most scenarios when more than four to five WL-classes were used indicates that the exact number of WL-classes does not significantly impact model calibration if at least four to five WL-classes are used. It also suggests that model performance should not be impacted dramatically if citizen scientists do not perfectly place the virtual staff gauge in the CrowdWater app so that the water level fluctuations do not cover all classes, as long as the water level fluctuates over at least four classes. In some cases, even fewer classes might be sufficient, especially if there is on average more than one observation per week, which is not unlikely when dedicated volunteers submit observations (Lowry et al., 2019).

More observations at higher flows and therefore higher WL-classes improved the model performance. Based on the significantly worse model performance for the *Crowd52* scenarios for which the percentage of observations during baseflow conditions was larger than 50% compared to scenarios for which this was less, we conclude that it is beneficial to encourage observations during times with larger water level fluctuations. This finding was also supported by the better model performance for the model calibrated with the *Crowd52* data for wet or average years compared to dry years. This is also the case when physical staff gauges (instead of virtual staff gauges for WL-class observations) are used. For some streams, particularly those with a flashy response, it may be difficult to get observations at high flow conditions because people are less likely to be outside and willing to stop to submit an observation. For other streams, this is possible, particularly when dedicated volunteers contribute regular observations because the high water levels are also very interesting for them (see example in Figure 3). A larger number of observations during these high flow periods can be obtained by sending push-messages if there is an app, text messages, e-mails, or social media posts.

Although the differences in model validation performance for the discontinuous water level and WL-class data were in most cases not statistically significant (Figures 5 and 6), there are advantages and disadvantages for both methods. The advantage of a real staff gauge is that citizens who pass by a location of interest may notice the staff gauge and are more directly invited to participate in the project. With the virtual staff gauge approach, this is not the case for people who have not installed the app yet (or haven't looked at the map of existing observation sites). Signpost to encourage participation could be used to highlight the virtual staff gauge site but this requires additional effort (for the project administrators to install the sign and for the citizen scientists to first download the app). Another advantage of a physical staff gauge is that at locations where the streambed doesn't change, the water level observations could be transformed to streamflow once a rating curve is available for that location. This is also possible for the WL-class data but of course results in an upper and lower bound of the streamflow for each WL-class observation (Strobl et al., 2019a). When the riverbed changes drastically either a new (virtual) staff gauge needs to be “installed” or the time series need to be considered separately. In case the data are used for model calibration, the alternative (though less preferable option) might be to use different parameter sets for the different periods.

The advantage of the virtual staff gauge approach is that it is easily scalable because only the citizen scientist needs to be at the location to set up a station and no equipment, permission, and local maintenance are required (Seibert, Strobl, et al., 2019). Of course, the use of an app, text messages, or even paper forms and mailboxes can also be combined. From a data quality perspective, the advantage of a virtual staff gauge approach to collect WL-class data using an app (e.g., the CrowdWater app) is that observations can be stored offline if no cellular network connection is available and can be uploaded later. Furthermore, the observations come with a picture of the situation, which allows some form of checking the data quality and allows further analysis using image recognition techniques. This is also possible for water level observations at real staff gauges but when only text messages are sent, the recipient must trust the sender that the water level reading and the time of the observation are correct. However, in areas without access to smartphones or internet, text messages or even paper forms might be the only option.

## 5. Conclusions

We studied the potential value of WL-class data that can be collected in citizen science projects for hydrological model calibration. Such data will be irregular in time, affected by errors and less precise than water level data. Our findings show that citizen science approaches to collect water level data using virtual or physical staff gauges with a few classes or precise markings are a promising way to obtain useful data for hydrological modeling in data-scarce catchments.

The results from the synthetic data sets indicated that time series with on average one WL-class observation per week over a 1-year period provides valuable information for calibration of a lumped bucket type model if there are four or more classes. Typical errors in the WL-class estimates for citizen science projects (Strobl et al., 2019) did not impact model performance considerably. Although the validation performance of the model calibrated with synthetic WL-class data with realistic frequencies for citizen science projects was not as good as when streamflow data were used for calibration, the performance was comparable to calibration with data collected with water level loggers or physical staff gauges with precise markings. The WL-class observation approach has the advantage of being easier to implement and more scalable because it does not require any physical installations and, thus, no special equipment, permits, or maintenance.

**Acknowledgments**
We thank all citizens who participated in the surveys to determine the typical error in the WL-class observations, the Swiss Federal Office for the Environment for providing the streamflow and water level data, MeteoSwiss for the provision of the weather data, Maria Staudinger for compiling the HBV input files, Marc Vis for HBV-model support, and the UZH Science Cloud for the provision of the infrastructure for cloud computing. The CrowdWater project is funded by the Swiss National Science Foundation (project no. 163008). The streamflow and water level data used for this study were obtained from the Swiss Federal Office for the Environment (FOEN); the station numbers are given in Table 1. The weather data were obtained from MeteoSwiss. The data repository for this study (Etter et al., 2019) contains the streamflow data from the FOEN for the selected calibration and validation years, the water level data, and WL-class data for all error types and observation scenarios, the model input and output files, the table with the parameter ranges, and the R-scripts.

## References

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The accuracy of citizen science data: A quantitative review. *The Bulletin of the Ecological Society of America*, *98*(4), 278–290. https://doi.org/10.1002/bes2.1336

Andréassian, V., Bourgin, F., Oudin, L., Mathevet, T., Perrin, C., Lerat, J., et al. (2014). Seeking genericity in the selection of parameter sets: Impact on hydrological model efficiency. *Water Resources Research*, *50*(10), 8356–8366. https://doi.org/10.1002/2013WR014761

Aschwanden, H., & Weingartner, R. (1985). *Die Abflussregimes der Schweiz. Publikation Gewässerkunde* (Vol, (Vol. 65). Bern, Switzerland: Geographisches Institut der Universität Bern, Abteilung Physikalische Geographie, Gewässerkunde. 1985

Balbo, A., & Galimberti, G. (2016). Citizen hydrology in River Contracts for water management and people engagement at basin scale. In *COWM2016 - International Conference on Citizen Observatories for Water Management*. Venice.

Bergström, S. (1976). Development and application of a conceptual runoff model for Scandinavian catchments. (S. Bergström, Ed.), *SMHI Rapporter* (Vol. 52). Norrköping, Sweden: SMHI Norrköping, Report RH07.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. Florence: Istituto superiore di scienze economiche e commerciali.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., et al. (2014). Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, *2*(26), 21. https://doi.org/10.3389/feart.2014.00026

Christophersen, N., Neal, C., & Hooper, R. P. (1993). Modelling the hydrochemistry of catchments: A challenge for the scientific method. *Journal of Hydrology*, *152*(1–4), 1–12. https://doi.org/10.1016/0022-1694(93)90138-Y

Cohn, J. P. (2008). Citizen Science: Can volunteers do real research? *Bioscience*, *58*(3), 192–197. https://doi.org/10.1641/B580303

Davids, J. C., van de Giesen, N., & Rutten, M. (2017). Continuity vs. the crowd—Tradeoffs between continuous and intermittent citizen hydrology streamflow observations. *Environmental Management*, *60*(1), 12–29. https://doi.org/10.1007/s00267-017-0872-x

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., et al. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, *10*(6), 291–297. https://doi.org/10.1890/110236

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, *30*(1), 192–197. https://doi.org/10.1214/aoms/1177706374

Etter, S., Strobl, B., Seibert, J., & van Meerveld, I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, *22*, 5243–5257. https://doi.org/https://doi.org/10.5194/hess-22-5243-2018

Etter, S., Strobl, B., Seibert, J., & van Meerveld, I. (2019). Data and R-scripts for: Value of crowd-based water level class observations for hydrological model calibration. https://doi.org/10.5281/zenodo.3371308

Fekete, B. M., Looser, U., Pietroniro, A., & Robarts, R. D. (2012). Rationale for monitoring discharge on the ground. *Journal of Hydrometeorology*, *13*(6), 1977–1986. https://doi.org/10.1175/JHM-D-11-0126.1

Hundecha, Y., Zehe, E., & Bárdossy, A. (2002). Regional parameter estimation from catchment properties prediction in ungauged basins. *Predictions in Ungauged Basins: PUB Kick-Off*, (December).

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S., & Burlando, P. (2011). The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model. *Water Resources Research*, *47*. https://doi.org/10.1029/2010WR009824

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., et al. (2018). Testing the waters: Mobile apps for crowdsourced streamflow data. *Eos*, *99*. https://doi.org/10.1029/2018EO096355

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.1436

Kundzewicz, Z. W. (1997). Water resources for sustainable development. *Hydrological Sciences Journal*, *42*(4), 467–480. https://doi.org/10.1080/02626669709492047

Lanfranchi, V., Wrigley, S., Ireson, N., Wehn, U., & Ciravegna, F. . (2014). Citizens' observatories for situation awareness in flooding. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May), 145–154.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing hydrologic data and engaging citizen scientists. *Ground Water*, *51*(1), 151–156. https://doi.org/10.1111/j.1745-6584.2012.00956.x

Lowry, C. S., Fienen, M. N., Hall, D. M., Stepenuck, K. F., & Paul, J. D. (2019). Growing pains of crowdsourced stream stage monitoring using mobile phones: The development of CrowdHydrology. *Frontiers in Earth Science*, *7*(128), 1–10. https://doi.org/10.3389/feart.2019.00128

McGuinness, J., & Bordne, E. (1972). *A comparison of lysimeter-derived potential evapotranspiration with computed values*. Washington, DC: Agricultural Research Service - United States Department of Agriculture.

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Meehan, T. D., Michel, N. L., & Rue, H. (2019). Spatial modeling of Audubon Christmas Bird Counts reveals fine-scale patterns and drivers of relative abundance trends. *Ecosphere*, *10*(4), e02707. https://doi.org/10.1002/ecs2.2707

Mulligan, M. (2013). WaterWorld: A self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrology Research*, *44*(5), 748. https://doi.org/10.2166/nh.2012.217

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I –A discussion of principles. *Journal of Hydrology*, *10*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Njue, N., Stenfert Kroese, J., Gräf, J., Jacobs, S. R., Weeser, B., Breuer, L., & Rufino, M. C. (2019). Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects. *Science of the Total Environment*, *693*, 133531. https://doi.org/10.1016/j.scitotenv.2019.07.337

Pool, S., Viviroli, D., & Seibert, J. (2019). Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States. *Water Resources Research*, *55*(1), 363–377. https://doi.org/10.1029/2018WR023855

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J., & McDonnell, J. J. (2015). Gauging the ungauged basin: Relative value of soft and hard data. *Journal of Hydrologic Engineering*, *20*(1), A4014004-1–6. https://doi.org/10.1061/(ASCE)HE.1943-5584.000086

Seibert, J., Strobl, B., Etter, S., Hummer, P., & van Meerveld, H. J. I. (2019). Virtual staff gauges for crowd-based stream level observations. *Frontiers in Earth Science*, *7*. https://doi.org/10.3389/feart.2019.00070

Seibert, J., van Meerveld, H. J., Etter, S., Strobl, B., Assendelft, R., & Hummer, P. (2019). Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen? *Hydrologie und Wasserbewirtschaftung*, *63*(2). https://doi.org/10.5675/HyWa_2019.2_1

Seibert, J., & Vis, M. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrological Processes*, *30*(14), 2498–2508. https://doi.org/10.1002/hyp.10887

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Sevruk, B. (1985). Systematischer Niederschlagsmessfehler in der Schweiz. In B. Sevruk (Ed.), *Der Niederschlag in der Schweiz* (pp. 31–65–75). Bern, Switzerland: Geographischer Verlag Kümmerly + Frey.

Show, H. (2015). Rise of the citizen scientist. *Nature*, *524*(7565), 265–265. https://doi.org/10.1038/524265a

Sideris, I. V., Gabella, M., Erdin, R., & Germann, U. (2014). Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society*, *140*(680), 1097–1111. https://doi.org/10.1002/qj.2188

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019a). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrological Sciences Journal*, 1–19. https://doi.org/10.1080/02626667.2019.1578966

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019b). The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. *PLoS ONE*, *14*(9), e0222579. https://doi.org/10.1371/journal.pone.0222579

Sy, B., Frischknecht, C., Dao, H., Consuegra, D., & Giuliani, G. (2018). Flood hazard assessment and the role of citizen science. *Journal of Flood Risk Management*, *12*(S2), e12519. https://doi.org/10.1111/jfr3.12519

van Meerveld, H. J., Vis, M. J. P., & Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, *21*(9), 4895–4905. https://doi.org/10.5194/hess-21-4895-2017

Weeser, B., Jacobs, S., Kraft, P., Rufino, M. C., & Breuer, L. (2019). Rainfall-Runoff modeling using crowdsourced Water level data. *Water Resources Research*, *55*. https://doi.org/10.1029/2019WR025248

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., et al. (2018). Citizen science pioneers in Kenya—A crowdsourced approach for hydrological monitoring. *Science of the Total Environment*, *631-632*, 1590–1599. https://doi.org/10.1016/j.scitotenv.2018.03.130